Kaplan and Huang *Large-scale Assess Educ* (2021) 9:15
https://doi.org/10.1186/s40536-021-00108-2

Large-scale Assessments in Education

METHODOLOGY Open Access

Bayesian probabilistic forecasting with large-scale educational trend data: a case study using NAEP

David Kaplan* and Mingya Huang

*Correspondence:
david.kaplan@wisc.edu
University of Wisconsin, Madison, USA

Abstract

Of critical importance to education policy is monitoring trends in education outcomes over time. In the United States, the National Assessment of Educational Progress (NAEP) has provided long-term trend data since 1970; at the state/jurisdiction level, NAEP has provided long-term trend data since 1996. In addition to the national NAEP, all 50 states and United States jurisdictions participate in the state NAEP assessment. Thus, NAEP provides important monitoring and forecasting information regarding population-level academic performance of relevance to national and international goals. However, an inspection of NAEP trend reports shows that relatively simple trend plots are provided; and although these plots are important for communicating general trend information, we argue that much more useful information can be obtained by adopting a Bayesian probabilistic forecasting point of view. The purpose of this paper is to provide a Bayesian probabilistic forecasting workflow that can be used with large-scale assessment trend data generally, and to demonstrate that workflow with an application to the state NAEP assessments.

Introduction

Of critical importance to education policy is monitoring trends in education outcomes over time. For example, on the international level, the United Nations Sustainable Development Goals identified Goal 4 as focusing on quality education for all, with Goal 4.6 focusing specifically on achieving literacy and numeracy for men and women (UN General Assembly, 2015). International large-scale assessment programs such as the OECD sponsored Program for International Student Assessment (PISA) (OECD, 2001) and IEA sponsored assessments in mathematics and science (TIMSS) and reading (PIRLS) (Mullis, 2013) can provide information to be used to forecast movement toward the goals set by the United Nations.

For the United States, the National Assessment of Educational Progress (NAEP) (US Department of Education, 2019) has provided long-term trend data since 1970; at the state level, NAEP has provided long-term trend data since 1996, and particularly after 2001 with the reauthorization of the Elementary and Secondary Education Act. In addition to the national NAEP, all 50 states and jurisdictions participate in the *State NAEP*

Assessment. Thus, throughout its history, NAEP has provided critical monitoring and forecasting information regarding trends in United States population-level academic performance.

Issues in trend reporting

A recent chapter by Kaplan and Jude (in press) provided an overview of trend analyses and reporting with international large-scale assessments. They argued that an inspection of trend reporting for PISA, PIRLS, and TIMSS reveals informative but relatively simple displays of changes in averages or percentages across time for populations and subpopulations of interest. The same holds true for NAEP; and although these displays are important for communicating trends to stakeholders, Kaplan and Jude argued that more detail could be gleaned from trend data by adopting a predictive model-based view of changes in trends over time. Kaplan and Jude (in press) further argued that a predictive model-based view of changes in trends over time could lead to the development of forecasting models which could supplement discussions of how countries are moving toward (or away from) nationally or internationally agreed-upon aims such as the UN Sustainable Development Goals.

The significance of adopting a predictive model-based view of trend analysis is twofold. First, this viewpoint can advance the policy and educational monitoring purposes of large-scale assessments. With respect to international large-scale assessments (ILSAs), a recent paper by Braun and Singer (2019) pointed out the problems associated with common uses of ILSAs. In particular, Braun and Singer (2019, p. 82.) noted that the use of ILSAs for evaluating curricular, instructional, and/or educational policies could be conducted but only with "extreme caution" and that using ILSAs for causal inference was "generally impossible". Braun and Singer (2019) did note however, that ILSAs were particularly useful for purposes of "transparency", to "...spur educational reforms" (e.g. the German "PISA shock"), to "describe and compare student achievement and contextual factors..." (with caveats), and, of relevance to this paper, "[t]o track changes over time" (again with some caveats). We agree with many of the issues raised in Braun and Singer (2019) and argue that ILSAs or national LSAs such as NAEP have not been sufficiently leveraged for one of the major purposes for which they were originally intended—namely, monitoring population-level trends in educational achievement. The predictive model-based framework that we are proposing in this paper can demonstrate the richness of policy information that can be obtained when using Bayesian prediction models to study educational trends at the population level.

Second, as described in more detail below, adopting a Bayesian framework allows us to directly address uncertainty in the parameters of our models and the models themselves. Directly addressing uncertainty has the benefit of yielding models that are known to possess optimal long-run predictive properties, and therefore should be preferred to more conventional frequentist methods when the goal is policy analysis.

We situate our predictive model-based approach within similar work conducted in economics looking at cross-national trends in economic growth (see Fernández et al., 2001c). First, perhaps obviously, we recognize that data must be longitudinal in order to study changes in trends over time. Clearly, NAEP data are longitudinal at the state level and thus, across states, constitute a panel data structure. Second, we follow the work of

Fernández et al. (2001c) by advocating for an approach toward forecasting that accounts for uncertainty in the parameters of change over time by implementing a fully Bayesian methodology (e.g, Gelman et al., 2014; Kaplan, 2014). Third, we argue along with Fernández et al. (2001c) that to be policy-relevant, it is necessary to identify predictors of change over time in educational outcomes of interest while at the same time recognizing the uncertainty in choosing any specific set of predictors as the "true" predictor set. Recognizing and accounting for the uncertainty in the selection of a forecasting model is also handled in a fully Bayesian framework.

The organization of this paper is as follows. In the next section, we provide a brief introduction to ideas in Bayesian statistical inference that are needed to set the notation and ideas that follow. Next, we discuss Bayesian growth curve modeling as the means by which we estimate states' growth rates accounting for uncertainty in the estimation of growth parameters. This is followed by a detailed discussion of Bayesian model averaging which we employ to account for uncertainty in the model used to predict the growth rates. Our empirical example using state data from NAEP follows.[1] We then turn to the results of the empirical example. This is then followed by a detailed sensitivity analysis of the findings, concentrating on the impact of different model and parameter priors on measures of predictive accuracy. The paper closes with a summary and conclusion focusing on assumptions, alternative methods, policy implications, and directions for future research.

Preliminaries on Bayes theorem and prior distributions

In this section, we set the notation and concepts of Bayesian statistics that will be necessary for later developments. Much more thorough treatments of Bayesian statistics can be found in Kaplan (2014) and Gelman et al. (2014).

The goal of statistical inference is to obtain estimates of the unknown parameters, denoted as θ.[2] The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of θ. In the frequentist tradition, the assumption is that θ is unknown, but has a fixed value that we wish to estimate. Measures such as the standard error or the frequentist confidence interval provide an assessment of the uncertainty associated with hypothetical repeated sampling from a population. In Bayesian statistical inference, θ is also considered unknown, however, similar to the data, θ is viewed as a random variable possessing a *prior probability distribution* that encodes our subjective or *epistemic* uncertainty (Howson & Urbach, 2006) about the true value of θ before having seen the data. Because both the observed data y and the parameters θ are assumed to be random variables, the probability calculus allows us to model the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution, namely

$$p(\theta, y) = p(y|\theta)p(\theta). \tag{1}$$

[1] The District of Columbia is a US Jurisdiction that participates in NAEP. From here on, we simply refer to states and jurisdictions as states.

[2] We note that θ could be a scalar value, such as a mean, or vector-valued, such as a set of regression coefficients.

where $p(\theta, y)$ is the joint distribution of the parameters and the data, $p(y|\theta)$ is the distribution of the data conditional on the parameters—i.e. the model, and $p(\theta)$ is the prior distribution. Bayes' theorem (Bayes, 1763; Laplace, 1774/1951) is then defined as

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{2}$$

where $p(\theta|y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data y representing our updated knowledge about the parameters of interest after having encountered the data and is equal to the data distribution $p(y|\theta)$ times the prior distribution of the parameters $p(\theta)$ normalized by $p(y)$ so that the posterior distribution sums (or integrates) to one.

Prior distributions

The general approach to considering the choice of a prior distribution on θ is based on how much information we believe we have *prior* to data collection and how accurate we believe that information to be. The strength of Bayesian inference lies precisely in its ability to incorporate our uncertainty about θ directly into our statistical models.

Non-informative priors

In some cases we may not be in possession of enough prior information to aid in drawing posterior inferences. Or, from a policy perspective, it may be prudent to refrain from providing subjective probabilities of effects of interest and instead, let the data speak for itself. Regardless, from a Bayesian perspective, this lack of information is still important to consider and incorporate into our statistical models. In other words, "...it is as equally important to quantify our ignorance as it is to quantify our cumulative understanding of a problem at hand" (Kaplan 2014, p. 18).

The standard approach to quantifying our ignorance is to incorporate non-informative prior distributions into our specification. In the case in which there is no prior knowledge to draw from, perhaps the most extreme non-informative prior distribution that can be used is the uniform distribution from $-\infty$ to $+\infty$, denoted as $\mathrm{U}(-\infty, +\infty)$ The uniform distribution essentially signals that we believe that our parameter of interest can take on an infinite number of values, each of which is equally likely. The problem with this particular specification of the uniform prior is that it is not proper insofar as the distribution does not integrate to 1.0. However, this does not always lead to problems, and is more of a conceptual issue. Highly diffused priors such as the $N(0, 10)$ distribution could also be used.

Weakly informative priors

Situated between non-informative and informative priors are *weakly informative* priors. Weakly informative priors are distributions that provide one with a method for incorporating less information than one actually has in a particular situation. Specifying weakly informative priors can be useful for many reasons. First, it is doubtful that one has complete ignorance of a problem for which a non-informative prior such as the uniform distribution is appropriate. Rather, it is likely that one can consider a more reasonable bound on the uniform prior, but without committing to much more information about

the parameter. Second, weakly informative priors are very useful in stabilizing the estimates of a model, particularly in cases of small sample sizes (see, Gelman et al., 2008). Specifically, Bayesian inference can be computationally demanding, particularly for hierarchical models, and so although one may have information about, say, higher level variance terms, such terms may not be substantively important, and/or they may be difficult to estimate, especially in small samples. Therefore, providing weakly informative prior information may help stabilize the analysis without impacting inferences.

Informative priors

Finally, it may be the case on the basis of previous research, expert opinion, or both, that information can be brought to bear on a problem and be systematically incorporated into the prior distribution. Such priors are referred to as *informative*. Informative prior distributions require that the analyst commit to the shape of the distribution. For example, if a parameter of interest, such as a regression coefficient is assumed to have a normal prior distribution, then the analyst must commit to specifying the average value and the precision around that value. Given that informative priors are inherently subjective in nature, they can be quite incorrect. Fortunately, Bayesian theory provides numerous methods for assessing the sensitivity of results to the choice of prior distributions. We will address sensitivity to choices of priors in our analysis below.

Bayesian v. frequentist comparisons

It is beyond the scope and purpose of this paper to outline all of the differences between Bayesian and frequentist methods, but several important distinctions relevant to this paper should be noted.

1. Bayesian inference is the only paradigm of statistics that allows for the quantification of subjective uncertainty. This form of uncertainty is not only present in our knowledge of the parameters of interest, but also in the very models that are used to estimate those parameters. Central to Bayesian theory and practice is that the intervals around parameter estimates (so-called *credible intervals*) are more accurate and models are more predictive if subjective uncertainty is directly addressed rather than ignored (Kaplan, 2014). We address both parameter and model uncertainty in this paper.
2. In the specific case of model uncertainty, Bayesian approaches can be shown theoretically to lead to optimal predictive models under specific assumptions (Clyde and Iversen, 2013, see also; Kaplan, 2021). Given that the goal of this paper is to describe an approach that can be used to develop optimally predictive models, we argue that the Bayesian approach is preferable to frequentist methods which do not incorporate subjective uncertainty nor have been shown to be superior to Bayesian methods in terms of optimal predictive performance. It should be noted, however, that in any single predictive analysis, frequentist methods might be better; nevertheless, Bayesian approaches have better long-run predictive performance on the basis of so-called *probabilistic scoring rules*. We will examine scoring rules for a variety forecasting models that vary in terms of their initial conditions.

3. In large samples, Bayesian approaches and frequentist approaches will converge, though their interpretations are different. As noted above, frequentist parameters are treated as fixed and only uncertainty due to sampling variability can be estimated through reference to the estimate's standard error. Bayesian estimates are interpreted probabilistically, and this, arguably, provides a much richer interpretation than the simple decision of whether a parameter estimate is statistically significant or not. We will highlight how Bayesian estimates provide interesting probabilistic interpretations as we proceed through the results.

Bayesian growth curve modeling

Our approach to probabilistic forecasting rests on the use of latent growth curve modeling (e.g., Bollen and Curran, 2006; Kaplan, 2009) wherein we treat the individual states as the units of analysis and estimate the trajectories in educational outcomes over time. We argue that latent growth curve modeling provides a flexible framework for estimating linear and non-linear trajectories and for incorporating predictors of the rate of change in academic outcomes. In a related context, the use of latent growth curve modeling for ex-post forecasting has been discussed and demonstrated in Kaplan and George (1998).

We situate our paper within the framework of linear growth curve modeling from the multilevel modeling perspective (see e.g., Raudenbush and Bryk, 2002). We write the intra-state (level-1) model as

$$y_{it} = \pi_{0i} + \pi_{1i} a_i + r_{ti} \tag{3}$$

where y_{it} is the outcome for state i ($i = 1, \ldots N$) at time t ($t = 1, \ldots T$), π_{0i} is the intercept capturing state i's status on the outcome at time t, π_{1i} is the slope (rate of linear growth over time) for state i, and r_{it} is the residual term. The term a_i marks the assessment cycles for state i. For this study, we use eight assessment cycles from 2003 to 2017. These are coded as $a_t = 0, 2, 4 \ldots 14$ to reflect that the cycles were every 2 years apart. This coding sets the intercept π_{0i} to be the math achievement score for state i in 2003. Together π_0 and π_1 are referred to as *growth parameters*.

The model in Eq. (3) is flexible enough to allow the growth parameters to be predicted by state-level time-invariant covariates. The inter-state (level-2) model can be written as

$$\pi_{si} = \beta_{s0} + \sum_{q=1}^{Q_s} \beta_{sq} x_{qi} + \epsilon_{si}, \tag{4}$$

where the π_{si} are the growth parameters (intercept and growth rate), x_{qi} are values on Q predictors for state i, β_{sq} are the regression coefficients, and ϵ_{si} are errors.

Latent growth curve modeling is very flexible. One important flexibility of latent growth curve modeling is the ability to incorporate time-varying covariates. These are variables that track the outcome of interest over time and are predictive of time-specific variation in the outcome. As will be discussed below, our focus will be on time-invariant outcomes because we are interested in modeling the average growth rate across states and not time-varying features of the growth trajectories.

Another important flexibility in latent variable growth curve modeling allows estimation of non-linear trajectories using *latent basis* methods. This specification requires that some of the time points in a_i be fixed to constants (e.g. 0, 2, 4) while allowing the remaining time points to be estimated from the data. Latent basis modeling yields data-based estimates of the time points and provides much better fit of the model to the empirical growth trajectories. We will apply latent basis methods in our example. These, and other extensions, are discussed in Bollen and Curran (2006).

For this paper, we will select a set of policyrelevant predictors for the latent growth curve model. Given that these predictors can change over time along with changes in mathematics achievement, we will form difference scores, subtracting 2017 values from 2003 values for each state and treat these difference scores as time-invariant predictors. We recognize that this is not an optimal solution, but present theory of BMA does not appear to provide for time-varying predictors in longitudinal models when the focus is on a growth parameter such as the linear growth rate.

A Bayesian framework for the growth curve model in Eqs. (3) and (4) first requires a specification of the probability model for the outcome—in our case the NAEP mathematics achievement scores. For our study, we will assume a normal probability model for mathematics achievement. Next, the Bayesian framework requires placing prior probability distributions on all model parameters (Kaplan, 2014). The choice of priors for our growth curve models are non-informative or weakly-informative (see e.g., Gelman et al., 2017) and will be discussed in more detail below.

Bayesian model averaging

An important aim of probabilistic forecasting is to identify policy-relevant predictors of growth. However, when confronting the problem of constructing probabilistic forecasting models, it is common to specify and estimate a set of different forecasting models and to use various model selection methods such as Akaike's information criterion (AIC) (Akaike, 1985, 1987) or the Bayesian information criterion (BIC) (Kass and Raftery, 1995; Schwarz, 1978) to choose a final model to report. The problem with model selection methods is that the analyst often proceeds as though the final selected forecasting model was the one considered in advance, and thus the uncertainty in the model selection process is ignored. More specifically, the selection of a particular model from a universe of possible models can be characterized as a problem of choice under pervasive uncertainty. The problem of model uncertainty has been nicely characterized by Draper and his colleagues who write (Draper et al., 1987, p. iii):

> *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 in Draper et al. (1987).]*

To address the problem of model uncertainty, we propose to use the method of *Bayesian model averaging* (Hoeting et al., 1999; Leamer, 1978; Madigan and Raftery, 1994; Raftery et al., 1997). The essential idea of Bayesian modeling averaging (described in more detail below) is to recognize that selecting a single model out of a class of possible models that could have been selected effectively ignores the uncertainty inherent in model choice. Model averaging diminishes this uncertainty by

taking a weighted average over a selected set of models, with weights that account for the quality of the model. These weights are referred to as posterior model probabilities (PMPs). Of importance to this paper is that theory and applications of Bayesian model averaging have shown that it provides better long-run predictive performance to that of any single model based on a class of scoring rules used in probabilistic forecasting analysis (Raftery et al., 1997). Specifically, BMA-based models are better calibrated to actual outcomes than any single model that one would choose, and the principle of choosing well-calibrated forecasting models is a key goal of predictive modeling (Dawid, 1982). Below, we will employ two specific scoring rules that we will use to evaluate the predictive performance of our forecasting models.

Bayesian model averaging has had a long history of theoretical developments and practical applications. Early work by Leamer (1978) laid the foundation for Bayesian model averaging. Fundamental theoretical work on Bayesian model averaging was conducted in the mid-1990s by Madigan and his colleagues (e.g., Hoeting et al., 1999; Madigan and Raftery, 1994; Raftery et al., 1997). Additional theoretical work was conducted by Clyde (1999). Draper (1995) has discussed how model uncertainty can arise even in the context of experimental designs, and Kass and Raftery (1995) provide a review of Bayesian model averaging and the costs of ignoring model uncertainty. A recent review of the general problem of model uncertainty along with applications to LSAs can be found in Kaplan (2021).

In addition to theoretical developments, Bayesian model averaging has been applied to a wide variety content domains. A perusal of the extant literature shows Bayesian model averaging applied to economics (e.g., Fernández et al., 2001b), bioinformatics of gene express (e.g., Yeung et al., 2005), weather forecasting (e.g., Sloughter et al., 2013), and propensity score analysis (Kaplan and Chen, 2014) to name just a few. An extension of Bayesian model averaging to structural equation modeling with applications to education can be found in Kaplan and Lee (2015) and a general review of Bayesian model averaging in the context of cross-sectional analyses of large-scale educational assessment data can be found in Kaplan and Lee (2018). A novel implementation of Bayesian model averaging to multiple imputation was proposed by Kaplan and Yavuz (2019)

This paper is strongly motivated by the work of Fernández et al. (2001b) who developed Bayesian model-averaged (BMA) growth regressions for gross domestic product over 140 countries from 1960 to 1992. Fernández et al. (2001b) found BMA-based growth regressions to be superior to any single model chosen on the basis of out-of-sample-predictive performance. In line with Fernández et al. (2001b), we propose to estimate a Bayesian latent growth regression model of NAEP 8th grade mathematics performance across the 50 states and the District of Columbia from 2003 to 2017. We then specify a set of predictors of growth and employ Bayesian model averaging to address model uncertainty. The end result will yield predictive densities of growth for each state allowing comparison of the actual growth rate in mathematics achievement and the growth rate predicted by the model. We will explore these models for 8th boys and girls separately.

Technical background of BMA

Following Madigan and Raftery (1994, see also; Kaplan, 2021), we will denote a future outcome of interest (in our case, state-level 8th mathematics achievement) as $\tilde{y}$. Next, consider a set of competing models $\{M_k\}_{k=1}^{K}$ that are not necessarily nested. The posterior distribution of $\tilde{y}$ given data y can be written as

$$p(\tilde{y}|y) = \sum_{k=1}^{K} p(\tilde{y}|M_k)p(M_k|y), \tag{5}$$

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

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l \neq k. \tag{6}$$

The important feature of Eq. (6) is that $p(M_k|y)$ will likely be different for different models, reflecting the relative uncertainty across models. The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{7}$$

where $p(\theta_k|M_k)$ is the prior distribution of the model parameters θ_k under model M_k (Raftery et al., 1997). Bayesian model averaging thus provides an approach for combining models specified by researchers, or perhaps elicited by key stakeholders.

Computational considerations for BMA

The number of models K appearing in the summation of Eq. (5) can be quite large. For example, if a regression model has $K = 10$ predictors, then there are $2^{10} = 1024$ models in the model space, M. The problem of reducing the overall number of models that one could incorporate in the summation of Eq. (5) has led to two interesting solutions. The approach that is used in the R software program BMS (Zeugner and Feldkircher, 2015) which we will employ in this paper is based on *Markov chain Monte Carlo Model composition* (MC^3)

Markov chain Monte Carlo model composition

The goal of Markov chain Monte Carlo model composition (MC^3) is to reduce the space of possible models that can be explored with Bayesian model averaging. Following Hoeting et al. (1999), the MC^3 algorithm proceeds as follows. First, let M represent the space of models of interest; in our case of growth curve modeling this would be the space of all possible combinations of variables used to predict the growth rates for each state. 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

$$\min\left\{1, \frac{pr(M'|y)}{pr(M|y)}\right\}, \tag{8}$$

otherwise, the chain stays in model M.

Sensitivity to parameter and model priors

One of the key steps when implementing BMA for probabilistic forecasting is to evaluate the predictive performance of BMA under different choices of parameter priors and model priors (Eicher et al., 2011; Feldkircher and Zeugner, 2009; Fernández et al., 2001a; Liang et al., 2008). Thus, for the last step in our analysis, we outline the methods we use for model evaluation and present results focusing on the sensitivity of our forecasting model utilizing choices that are readily available in the software program that we implemented for this study—BMS (Zeugner and Feldkircher, 2015).

Scoring rules for BMA

As noted earlier, a central characteristic of statistics is to develop accurate predictive models (Dawid, 1984). Indeed, as pointed out by Bernardo and Smith (2000), all other things being equal, a given model is to be preferred over other competing models if it provides better predictions of what actually occurred. Thus, a critical component in the development of accurate predictive models is to decide on rules for gauging predictive accuracy—often termed *scoring rules*. Scoring rules provide a measure of the accuracy of probabilistic forecasts, and a forecast can be said to be *well-calibrated* if the assigned probabilities of the outcome match the actual proportion of times that the outcome occurred.

Following closely the discussion in Kaplan (2021, pp. 222–226), a number of scoring rules have been proposed for probabilistic forecasting (see e.g., Bernardo and Smith, 2000; Gneiting and Raftery, 2007; Jose et al., 2008; Merkle and Steyvers, 2013; Winkler, 1996); however, for this paper we will primarily evaluate predictive performance under different parameter and model prior settings using the log predictive score (Bernardo and Smith, 2000; Good, 1952) and the Kullback–Leibler divergence (KLD) (Kullback, 1959, 1987; Kullback and Leibler, 1951).

The log predictive score (LPS) is defined as

$$-\sum_i \log\left[p(\tilde{y}_i|x, y, \tilde{x}_i)\right] \tag{9}$$

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

In addition to the log predictive score we also use the Kullback–Leibler divergence (KLD). The KLD between two distributions can be written as

$$\text{KLD}\,(f,g) = \int f(y) log\left(\frac{f(y)}{g(y|\theta)}\right) dx \tag{10}$$

where KLD (f,g) is the "information lost when g is used to approximate f. In our case, the estimated growth rate without predictors is compared to predicted growth rate using Bayesian model averaging along with different choices of model and parameter priors. The forecasting model with the lowest KLD measure is deemed best in the sense that the information lost when approximating the "true" outcome distribution with the distribution predicted on the basis of the model is lowest. For this paper, LPS values will be obtained using BMS, and the KLD values will be obtained using the R package LaplacesDemon (Statisticat, & LLC 2020).

Parameter priors

The choice of parameter priors available for our sensitivity analyses rest 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 for model, written as,

$$y_i = x_i'\beta + \varepsilon, \tag{11}$$

where ε is i.i.d. $N(0,\phi^{-1})$. For a give model, say M_k, Zellner's g-prior can be written as

$$\beta_k|\sigma^2, M_k, g \sim N\left(0, \sigma^2 g\left(x_k' x_k\right)^{-1}\right). \tag{12}$$

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

The g-prior is not without criticisms. In particular, Feldkircher and Zeugner (2009) have pointed out that prior choices of g can have very large impacts 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 (Eicher et al., 2011; Feldkircher and Zeugner, 2009; Fernández et al., 2001a; Liang et al., 2008) 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*. The list of fixed and flexible model priors used in this study are as follows (Zeugner and Feldkircher, 2015):

Fixed priors

- Unit information prior: $g = N$ (in our case $g = 50$). This prior was used in our empirical example. 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): $g = Q^2$ Foster and George (1994) show 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 the use of HQ = 3 for large N.

Flexible priors

- 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; Hansen and Yu, 2001; Liang et al., 2008).
- Hyper-g prior: 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). In this study, five types of hyper-g priors are examined in line with Liang et al. (2008) and Feldkircher and Zeugner (2009):

  - HG-3: Setting $\alpha = 3$ results in the prior for the shrinkage factor to be $\frac{2}{3}$.
  - HG-4: Setting $\alpha = 4$, results in an approximately uniform prior for the shrinkage factor.
  - HG-UIP: $\alpha = 2 + \frac{2}{n}$ yields the "g-UIP"-shrinkage where factor to be $E\left(\frac{g}{1+g}\right) = \frac{N}{1+N}$.
  - HG-RIC: $\alpha = 2 + \frac{2}{Q^2}$ yields the "g-RIC"-shrinkage factor where $E\left(\frac{g}{1+g}\right) = \frac{Q^2}{1+Q^2}$.
  - HG-BRIC: sets the prior shrinkage factor $E\left(\frac{g}{1+g}\right)$ to be equivalent to the BRIC.

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 12 candidate default priors and concluded that the unit information prior (UIP) combined with the uniform model prior outperformed the other choices. In the following section, we examine how BMA performs under Zellner's g-prior setting based on the LPS and KLD scores.

Model priors

For this paper, three model priors are investigated and 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 $\sum_{q=0}^{Q} \binom{Q}{q} q 2^{-Q} = Q/2$. So, for our analysis, the expected model size would be $6/2 = 3$. 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 $\theta < .5$ will weight the posterior mass toward smaller models, and visa versa (Zeugner and Feldkircher, 2015). For this study, the default prior model size for binomial model prior is $Q/2 = 6/2 = 3$ for this study. Therefore, to assess the impact of model prior size when using the binomial model prior, we choose one unit lower (model prior size = 2) and higher (model prior size = 4).
- 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. $\theta \sim$ Beta (a, b). Under the Beta-binomial prior the first and second moments of the model size $\bar{m}$ are,

$$E(\bar{m}) = \frac{a}{a+b}Q, \tag{13}$$

$$var(\bar{m}) = \frac{ab(a+b+Q)}{(a+b)^2(a+b+1)}Q \tag{14}$$

Empirical example: analysis of state NAEP data, 2003–2017

Two data sources were combined to provide the variables necessary for this analysis. First, a data file was constructed to provide the state NAEP mathematics achievement and reading data from 2003 to 2017, eight assessment cycles in total. Other data sources in the NAEP data file that were used for the predictive modeling (described in more detail below) are the National School Lunch Program (NSLP) variables obtained as the percentage of students in the state who are NSLP-eligible, and taken as a proxy for socio-economic status. In addition, demographic variables such as percentage of gender groups were also included in this data file.

The NAEP data file was merged with specific variables in the Common Core of Data (CCD) (2020) to obtain information regarding state staff counts in the metric of *full-time equivalents* (FTEs), per pupil state revenue, pupil/teacher ratio. It should be noted that Tennessee was excluded from this analysis due to lack of state reported data in the CCD file.

Analysis steps

In this section, we provide the analysis steps for estimating our Bayesian probabilistic forecasting model. We argue that these steps describe a reasonable workflow for probabilistic forecasting and may be general enough to employ in other relevant contexts.

1. We began by specifying a simple cross-state latent growth model of 8th grade mathematics achievement for states that participated in the state NAEP assessments since 2003.[3] This choice provided the most complete set of panel data possible on which to develop our models. We focused our analyses separately for boys and girls. This first step was necessary in order to determine the general shape of the achievement trends for the states over time. On inspection of the trajectories, we employed the flexibility of latent basis growth curve modeling discussed above, and determined the choice of the functional form of the model based on a number of model selection measures—particularly the Bayesian information criterion (Kass and Raftery, 1995).

   Throughout, we used the blavaan software package (Merkle and Rosseel, 2018). The blavaan software package uses the lavaan (Rosseel, 2012) syntax structure to call the Bayesian software program rjags (Plummer, 2014). The rjags software program implements Markov chain Monte Carlo (MCMC) sampling via the Gibbs sampler (see e.g., Gilks et al., 1996). Weak or non-informative priors for the Bayesian growth curve parameters are shown in the last column of Table 1. For this study, we used two chains with 1000 adaptation steps, 5000 burn-in steps, and 100,000 post burn-

[3] Note that state-level estimates published by National Center for Education Statistics (e.g., found in the *Nation's Report Card* (US Department of Education, 2019) or the *NAEP Data Explorer* (US Department of Education, 2021)) automatically incorporate plausible values and appropriate weighting. See https://nces.ed.gov/nationsreportcard/tdw/ for more details.

Table 1 Bayesian growth curve modeling results for boys and girls

	Estimate	Post.SD	HPD.025	HPD.975	PSRF	Prior
Boys growth params.						
Intercept	274.782	1.300	272.116	277.226	1.000	dnorm(260,.1)
Slope	0.931	0.108	0.726	1.149	1.000	dnorm(0,1e−2)
Pre(intercept)	88.460	19.415	55.216	127.007	1.000	dwish(iden,3)
Pre(slope)	0.246	0.067	0.132	0.385	1.000	dwish(iden,3)
Girls growth params.						
Intercept	273.707	1.308	271.113	276.219	1.000	dnorm(260,.1)
Slope	0.902	0.122	0.669	1.142	1.002	dnorm(0,1e−2)
Pre(intercept)	85.114	19.065	51.593	122.738	1.000	dwish(iden,3)
Pre(slope)	0.269	0.084	0.126	0.43	1.000	dwish(iden,3)

pre() refers to the precision of the parameter, where precision = 1/variance. *dnorm* is the normal $N(\mu, \tau^2)$ distribution, where τ^2 is the precision. *dwish* is the Wishart (R, ν) distribution for the precision matrix with shape matrix, R and degrees-of-freedom, ν

in iterations. To assess convergence of the chains, we inspected trace plots, density plots, and autocorrelation plots, and in all cases, the MCMC algorithm converged. In addition, we used the potential scale reduction factor (PSRF) (Brooks and Gelman, 1998; Gelman and Rubin, 1992) which calculates the ratio of the between-chain variance to the within-chain variance. Convergence of the chains is achieved when the PSRF is less than to 1.01. The R code used for this study is available on our website at http://bmer.wceruw.org/index.html.

2. In addition to the 8th grade mathematics achievement measures, we selected a set of variables deemed to be policy-relevant predictors of the shape of the trends over time. Our choice of predictors was guided by an inspection of the student, teacher, and school questionnaires and other data sources. Also for this step, we calculated difference scores between the 2017 and 2003 measures of these variables (if available) and used these change scores as predictors of growth in 8th mathematics achievement.
3. We next implemented Bayesian model averaging to our forecasting model and compared the results to the model estimated in Step 2, again on the basis of forecasting statistics. This was accomplished in a two-step process by computing the random growth parameters for each state and importing them into the Bayesian model averaging software. The results of Bayesian model averaging provide an assessment of the impact of the chosen predictors on growth, accounting for model and parameter uncertainty.

   For this step, we initially specified the default parameter and model priors available in the software program BMS (Zeugner and Feldkircher, 2015). For the priors on the model parameters we used the unit information prior discussed earlier (Raftery, 1998). The unit information prior is a weakly informative prior that is diffused over the region of the likelihood where parameter values are considered mostly plausible, but not overly spread out. This is accomplished by forming the prior based on the maximum likelihood estimate of the parameter mean, with variance equal to the expected information from one observation. The default prior on the model space is $1/M$, where M is the number of models, reflecting the belief that no model is to be

favored a priori as the true forecasting model. Other priors will be explored in the section of this paper on sensitivity analysis.

4. We next used the results from the Bayesian model averaging step to obtain posterior predictive densities of growth rates in 8th mathematics achievement for the boys and girls separately across the states. These predictive densities provide a means of checking the prediction model of the growth rate against the actual growth, allowing examination of problems with the prediction model and/or potential outlier states. Also, 95% quantiles around the predicted growth rates are provided.
5. Finally, we examined the sensitivity of our results to different choices of parameter and model priors, that serve as initial settings for BMA. We view this step as essential before utilizing any forecasting model for serious policy analysis.

Bayesian growth curve modeling results

In this section we present the results of the Bayesian growth curve model described earlier to fit the mathematics achievement trend data for 2003 thru 2017 and for boys and girls separately. Preliminary analyses suggested estimating the latent basis model, allowing data-based estimation of the time coefficients so that the slope corresponds to the unique shape of the trend. In particular, the first three time points (2003, 2005, and 2007) were set to fixed values while the remaining time points were estimated from the data (see Bollen and Curran, 2006, for a discussion of this method).

The Bayesian growth curve modeling results for boys and girls separately on NAEP 8th grade mathematics achievement are shown in Table 1. The "intercept" parameter is set to be the average mathematics achievement score in 2003. The posterior parameter estimates are given under the "Estimate" column along with their posterior standard deviations in the next column. This is followed by the 95% highest posterior density (HPD) interval.[4] The column labeled "PSRF" contains the potential scale reduction factors for each parameter and serves as an assessment of the convergence of the MCMC chains.

The final column shows the priors that were used for this analysis. For all but the intercept parameter, non-informative priors were used. Specifically, for the intercept and slope, a normal prior was used. For the intercept, a default value of zero led to MCMC convergence issues. Because it is known that the average NAEP mathematics achievement score for our groups is much greater than zero, we placed an informative prior on the intercept to help attain convergence. The precision on the normal prior for the intercept was set to .1 indicating relatively high certainty regarding the mean value. The precision for the normal prior on the slope was set much lower indicating relatively less certainty.[5] Priors used for the precisions of the intercept and slope were given Wishart distributions with non-informative shape and scale parameters (see Kaplan, 2014, for a discussion of these priors).

For boys in the top panel of Table 1, we note on the basis of the PSRF that all parameters converged to their posterior distributions. We find that the posterior estimate of

[4] The HPD is an interval in which every point inside the interval has a higher probability than any point outside the interval (Kaplan, 2014).

[5] Note that the variance of the prior is 1/precision, and so here the variance is 10.

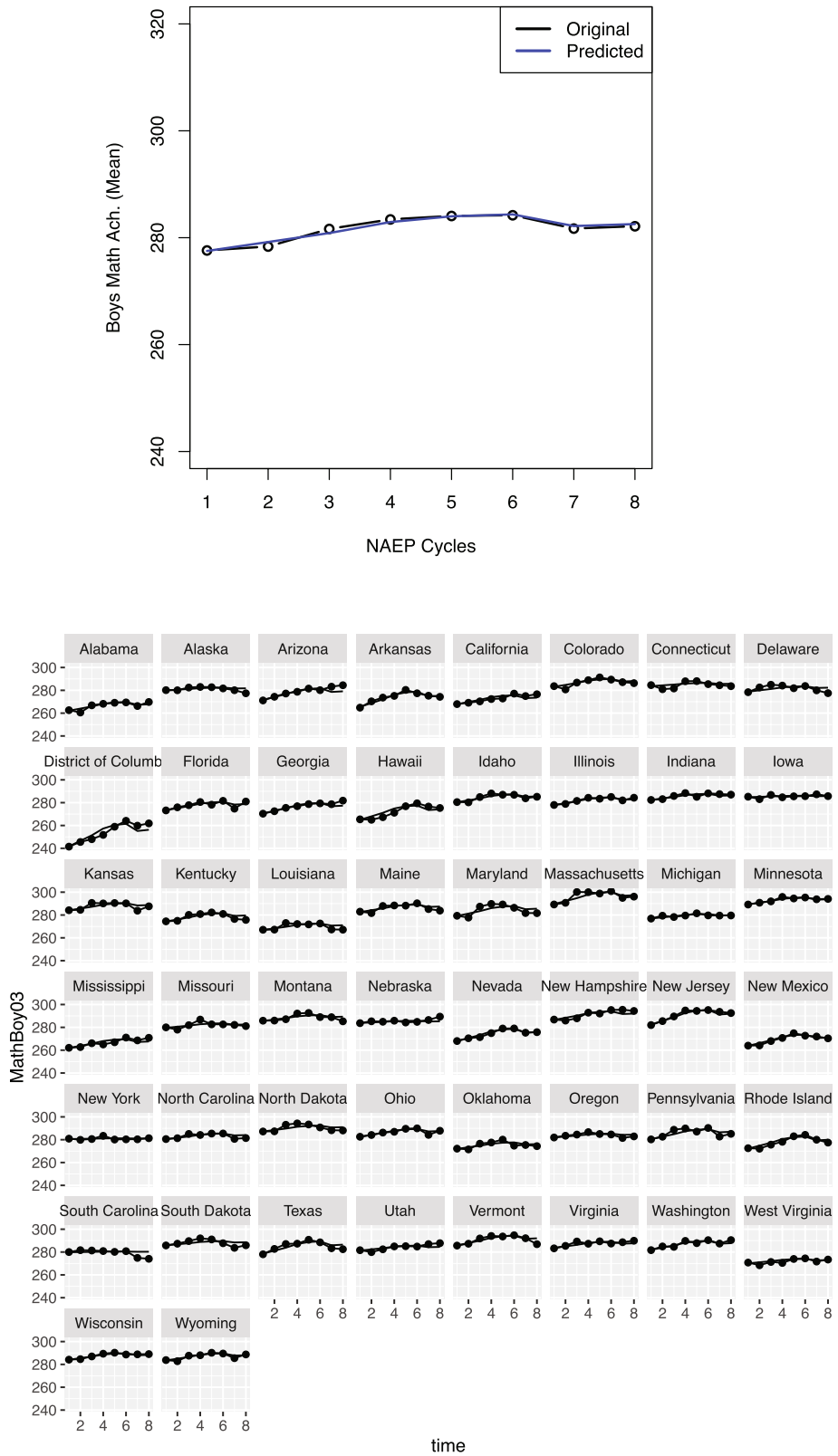

Fig. 1 Country-level (above) and state-level (below) fitted trends in 8th grade boys mathematics achievement

the rate of mathematics achievement growth for boys is 0.931 (sd = 0.108). The HPD indicates that there is a 95% probability that the true mathematics achievement growth for boys is between 0.726 and 1.149. In addition, we find that the probability that the effect is greater than zero is approximately 1.0. The country-level and state-level fitted and actual trends are displayed in Fig. 1.

The results for 8th grade girls' mathematics achievement are displayed in the lower panel of Table 1. Here too, we note that on the basis of the PSRF all parameters have converged to their posterior distributions. We find that the posterior estimate of the rate of mathematics achievement growth for girls is 0.902 (sd = 0.122). The HPD indicates that there is a 95% probability that the true mathematics achievement growth for girls is between 0.669 and 1.142 and the probability that the growth rate is greater than zero is approximately 1.0. The country-level and state-level fitted and actual trends for girls are displayed in Fig. 2.

Three important points regarding the growth curve modeling results should be highlighted. First, note that for both boys and girls we used latent basis modeling, and thus the fit of the model to the empirical trajectories is expected to be quite good. Second, although the model estimated rate of growth across the 50 states is almost certainly not zero, an inspection of the individual state trajectories in Fig. 1 show that they are quite small and the trends appear flat for most states. Finally, note that the description of the findings are stated in probabilistic terms. This includes statements about the highest probability density as well as statements about the probability that the estimates are greater than zero. Indeed, any probability statement of relevance to a research question is possible insofar as the posterior probability distribution of the parameter is available and can be summarized via simple statistical calculations. Such probability statements are not possible in the context of frequentist growth curve modeling, which would simply render a dichotomous decision about whether the average growth rate is significantly different from zero.

Bayesian model averaging results

In this section, we provide the results of using Bayesian model averaging over a set of selected predictors of growth in 8th mathematics achievement for boys and girls separately. The variables used in this model are

1. BoysEnrollDiff: The difference between the 2017 8th grade enrollment of boys and 2003 8th grade enrollment of boys expressed as a percentage ($\times 100$) of the population of 8th grade enrollment. The variable is used for the analysis of the boys' data.
2. GirlsEnrollDiff: The difference between the 2017 8th grade enrollment of girls and 2003 8th grade enrollment of girls expressed as a percentage 8th grade enrollment of boys and 2003 8th grade enrollment of boys expressed as a percentage ($\times 100$) of the population of 8th grade enrollment. This variable is used in the analysis of the girls' data.
3. PTRatioDiff: The difference between 2017 pupil/teacher ratio and the 2003 pupil/teacher ratio
4. FTEdiff2: The difference in the 2015 and 2003 state level full time equivalent teachers (divided by 10,000). Data were unavailable for 2017.

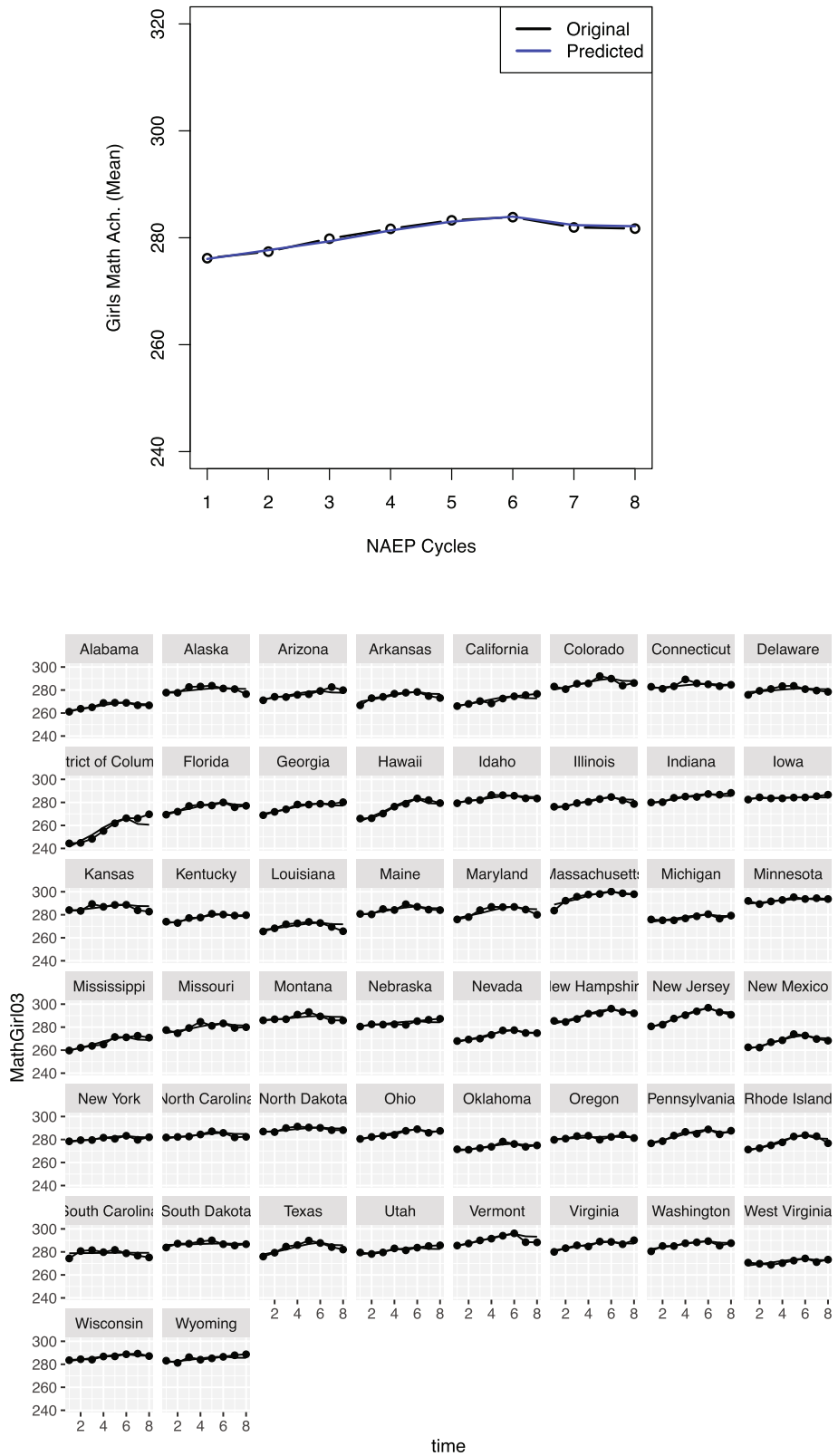

Fig. 2 Country-level (above) and state-level (below) fitted trends in 8th grade girls mathematics achievement

Table 2 Bayesian model averaging results for boys and girls

	PIP	Post Mean	Post SD	Cond. Pos. Sign
Boys				
ReadBoyDiff	1.00	0.05	0.01	1.00
TOTREVdiff	0.21	0.05	0.13	1.00
PTRatioDiff	0.19	− 0.01	0.03	0.00
BoysEnrollDiff	0.16	0.01	0.02	1.00
NSLPLunchDiff	0.12	0.00	0.00	0.87
FTEdiff2	0.12	− 0.00	0.00	0.37
Girls				
ReadGirlDiff	0.98	0.06	0.02	1.00
GirlsEnrollDiff	0.72	− 0.09	0.07	0.00
PTRatioDiff	0.46	− 0.04	0.05	0.00
TOTREVdiff	0.41	0.16	0.24	1.00
NSLPLunchDiff	0.13	− 0.00	0.00	0.28
FTEdiff2	0.13	0.00	0.00	0.95

PIP: Posterior inclusion probability; Post Mean: expected a posteriori estimate; Cond.Pos.Sign: Probability that the sign of the estimate is positive conditional on inclusion in the model

5. TOTREVdiff: The difference between the 2015 and 2003 total revenue (divided by 10,000)
6. NSLPLunchDiff: The difference in the 2017 and 2003 percentage of NSLP-eligible students
7. ReadDiff: The difference in 2017 and 2003 NAEP reading scores.

The outcome variable is the growth rate in 8th mathematics (for boys and girls separately) obtained from the Bayesian growth curve model. With six predictors for each model, there are $2^6 = 64$ possible models to be explored.[6] We use the R program BMS (Zeugner and Feldkircher, 2015) to explore the space of possible models, yielding weighted averaged regression coefficients, with weights corresponding to the posterior model probabilities (PMPs) of the models retained by the algorithm as described in Eq. (8).

The Bayesian model averaging results are shown in Table 2. The column labeled "PIP" shows the posterior inclusion probabilities for each variable, referring to the proportion of times the variable appeared in the models searched by the algorithm. For example, the PIP for ReadBoyDiff is 1.00, meaning that across all the models selected by the algorithm, the 2017–2003 difference in 8th grade reading for boys appears in 100% of the models. 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 conditional probability that the sign of the respective regression coefficient is positive conditional on its inclusion in the model. For example, with ReadBoyDiff, the probability is 1.0 that the sign is positive - that is, the coefficient's

[6] We note that in typical applications of BMA, many more predictors are used and thus a much larger model space is explored. We discuss this issue in the "Summary and conclusions" section.

sign is positive for all models explored by the algorithm. In contrast, for girls, the probability that the sign for NSLPLunchDiff is only 0.28, indicating a low probability that NSLPLunchDiff is positive.

For both boys and girls, we find that the 2017–2003 reading difference score is the strongest (model averaged) predictor of growth in mathematics. The sign of the posterior means for boys and girls are positive indicating that the positive change in reading over the years is associated with a positive change in the growth in mathematics achievement. In each case, the PIPs are very close to one and the effects are almost certainly positive. Remaining effects have appreciably smaller posterior inclusion probabilities, particularly for boys. For girls, however, we find that positive change 8th enrollment from 2003 to 2017 appears in 72% of the models explored by the algorithm and is associated with a positive rate of growth in mathematics achievement.

Figure 3 displays the posterior coefficient density plots for boys (upper panel) and girls (lower panel). These figures provide a visual display of the information in Table 2. We find, as expected under the model, that the posterior densities are normal and that the medians and mean values of the coefficients (COND EV) align very well. The dashed lines are the 95% posterior probability intervals.[7]

Bayesian model averaging also provides information on the posterior model probabilities (PMPs) of each model searched by the algorithm. Table 3 provides the PMPs for the top five models for boys and girls. Three important results should be noted. First, for boys and girls, the PMP for the top model (Model 1) is quite small (0.40 for boys and 0.17 for girls). In a typical model selection framework, this top model would be chosen, and consistent with the quote from Draper et al. (1987) referenced earlier, the inferences from this top model would be over-confident because of the considerable uncertainty in this model, and decisions based on predictions from this model would be very risky.[8] Second, and in a similar vein, the total posterior probability of the top 5 models is considerably less than 1.0, again suggesting extensive uncertainty across the space of models. As this is a case study with relatively few predictors, these results are not surprising, but do serve as a caution. Third, if we were to simply use all of the variables in a single model, we would still be assuming that the posterior model probability of that full model is 1.0, which it might not be, and thus again we would be in danger of drawing overconfident inferences.

Predictive densities

For the next step in our analysis, we provide Bayesian predictive densities of growth in 8th grade mathematics. As an example, we focus on predictive densities for boys and girls in Colorado and DC. These plots are shown in Fig. 4. Results for all remaining states are available as Additional file 1. The dashed vertical line is the actual growth rate and the solid line is the model-predicted growth rate based on Bayesian model averaging. The 95% quantiles for the predictive densities are also displayed. These figures provide a means of judging the adequacy of the model in terms of how predictions of growth

[7] We note that the PIP values displayed in Fig. 3 are not rounded as compared to Table 2.

[8] Note that the top model would also be associated with the lowest Bayesian information criterion, and would be selected on that basis as well.

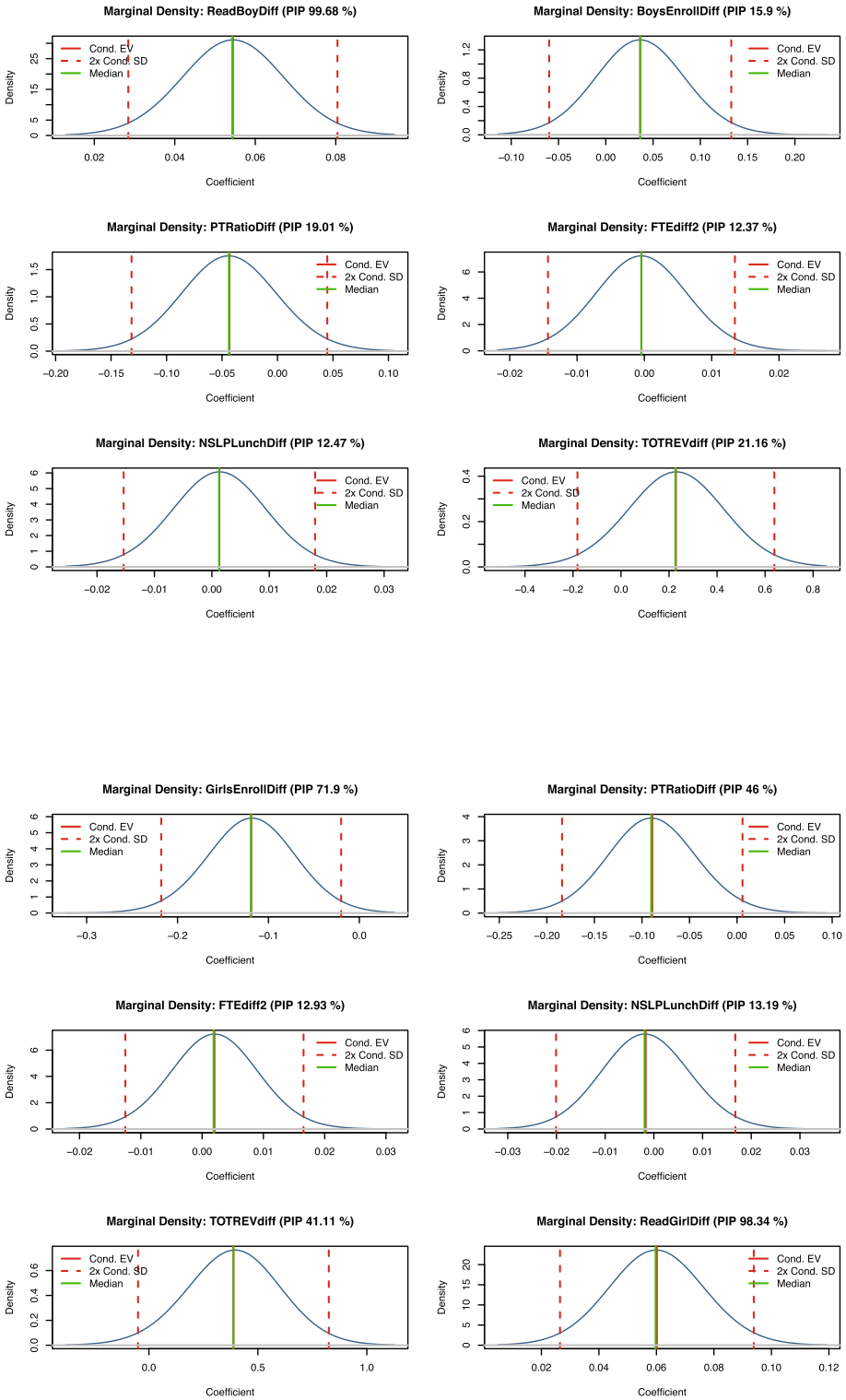

Fig. 3 Marginal posterior density plots and posterior inclusion probabilities for boys (top) and girls (bottom)

Table 3 Posterior model probabilities for top five models

	Model 1	Model 2	Model 3	Model 4	Model 5
PMP (Boys)	0.40	0.12	0.10	0.08	0.06
PMP (Girls)	0.17	0.16	0.14	0.09	0.06

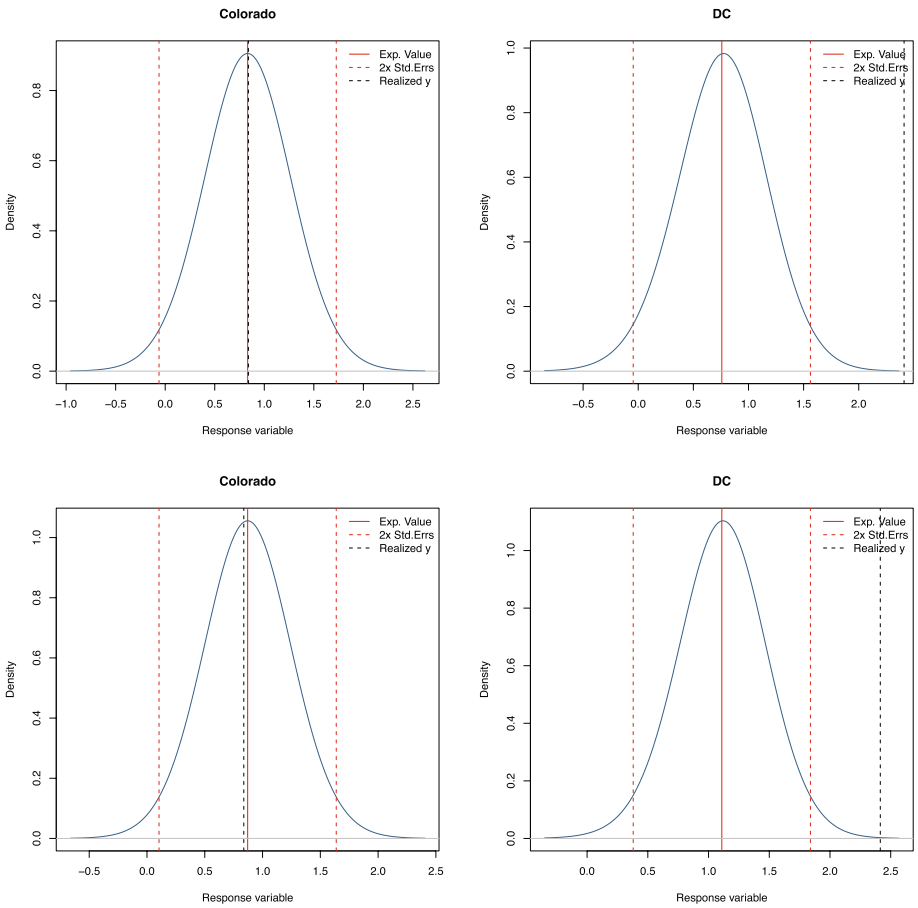

Fig. 4 Boys (top row) and girls (bottom row) posterior forecast density plots: Colorado and DC

align with the actual growth rate. Large differences between the observed and predicted growth rates are indicative of the prediction model being incorrect for that state, that the state may be an outlier, or both. An example of the prediction model being a good fit to the actual growth rate can be seen in Fig. 4 for boys' mathematics achievement, with the prediction of 8th grade growth in mathematics for Colorado. An instance in which the model is not predicting growth accurately is DC. In the situation of DC, it may be necessary to examine the data closely for possible problems, or examine different settings for

Table 4 Summary of log-predictive scores and Kullback–Leibler divergences for boys and girls in the fixed priors setting

	LPS				KLD			
	UIP	RIC	BRIC	HQ	UIP	RIC	BRIC	HQ
Boys								
Uniform	0.391	0.391	0.391	0.392	0.098	0.097	0.098	0.098
Binomial (m = 2)	0.391	0.396	0.391	0.397	0.098	0.098	0.098	0.098
Binomial (m = 4)	0.387	0.391	0.387	0.387	0.097	0.097	0.097	0.097
Beta-binomial	0.397	0.395	0.397	0.397	0.098	0.098	0.098	0.098
Girls								
Uniform	0.391	0.390	0.391	0.392	0.096	0.096	0.096	0.096
Binomial (m = 2)	0.401	0.402	0.401	0.405	0.097	0.096	0.097	0.097
Binomial (m = 4)	0.383	0.390	0.383	0.383	0.096	0.096	0.096	0.096
Beta-binomial	0.393	0.390	0.393	0.395	0.096	0.096	0.096	0.096

UIP: Unit information prior; RIC: Risk inflation criterion; BRIC: Benchmark risk inflation criterion; HQ: Hannan–Quinn criterion; Uniform: uniform model prior; Binomial (m = 2): Binomial model prior with model size = 2; Binomial (m = 4): Binomial model prior with model size = 4; Beta-binomial: Beta-binomial model prior

the prediction model, such as alternative priors placed on the model parameters. We examine overall predictive accuracy next. We observe that the results are virtually identical for girls.[9]

Sensitivity test results

The results of our sensitivity tests for boys and girls under the fixed priors setting are displayed in Table 4. We find that the LPS and KLD values within model prior settings are virtually identical across all parameter priors. However, across model prior settings, the binomial model prior with m = 4 yields a lower LPS compared to other the other model priors regardless of the choice of parameter priors. The KLD values for the analysis of the boys are slightly lower under the binomial prior with m = 4 setting while the KLD results for the girls is less clear.

The results for boys and girls under the flexible priors setting are found in Table 5. As with the fixed prior setting, the LPS and KLD values are also quite similar across different parameter priors setting. However, the LPS values in the flexible prior setting are slightly larger overall compared to the values in the fixed priors setting while KLD values are almost the same. In the uniform model prior condition, using the local empirical Bayes prior obtains the smallest LPS while using the hyper-g-prior with the $\alpha = 4$ yields the highest LPS values.[10] The KLD values remain mostly consistent over all flexible priors setting. Because $Q^2 = 36$ and $N = 50$ in this study, the hyper-*g*-BRIC prior translates the shrinkage factor to be $\arg\max(Q^2, N)$ to be N, which is identical to the hyper-*g*-UIP prior. In general, the models in the fixed priors setting yield smaller LPS values than the ones in the flexible priors setting while the KLD values are almost identical in both setting.

[9] There are other states where the model is not predicting the actual growth rate very well. Results for DC, however, were the most extreme.

[10] See however, the footnote in Table 5 for a caveat.

Table 5 Summary of log-predictive scores and Kullback–Leibler divergences for boys and girls in the flexible priors setting

	LPS						KLD					
	EBL	**HG-3**	**HG-4**	**HG-UIP**	**HG-RIC**	**HG-BRIC**	**EBL**	**HG-3**	**HG-4**	**HG-UIP**	**HG-RIC**	**HG-BRIC**
Boys												
Uniform	0.395*	0.402	0.410	0.398	0.397	0.398	0.098*	0.098	0.098	0.097	0.097	0.097
Binomial (m = 2)	0.397	0.404	0.411	0.403	0.401	0.403	0.097	0.098	0.098	0.098	0.097	0.098
Binomial (m = 4)	0.395*	0.403*	0.412*	0.397	0.395	0.397	0.098*	0.099*	0.099*	0.097	0.097	0.097
Beta-binomial	0.397	0.405	0.412	0.407	0.405	0.412	0.097	0.098	0.099	0.098	0.097	0.098
Girls												
Uniform	0.397*	0.406	0.412	0.401	0.400	0.401	0.095	0.096	0.097	0.096	0.096	0.096
Binomial (m = 2)	0.406	0.416	0.423	0.413	0.413	0.413	0.096	0.097	0.098	0.097	0.097	0.097
Binomial (m = 4)	0.392*	0.400*	0.407*	0.394	0.394	0.394	0.095*	0.096*	0.096*	0.095	0.095	0.095
Beta-binomial	0.394	0.403	0.410	0.401	0.394	0.401	0.096	0.096	0.097	0.096	0.096	0.096

Values with * indicate that these analyses was based on fewer than the full 64 total possible models due to computational tolerances being reached in the calculation of likelihoods. In no case were fewer than 30 models used in the LPS and KLD calculations for these cases

EBL: Local empirical Bayes; HG-3: Hyper-*g* prior with $\alpha = 3$; HG-4: Hyper-*g* prior with $\alpha = 4$; HG-UIP: Hyper-*g* prior with UIP setting; HG-RIC: Hyper-*g* prior with RIC setting; HG-BRIC: Hyper-*g* prior with BRIC setting. Uniform: uniform model prior; Binomial (m = 2): Binomial model prior with model size = 2; Binomial (m = 4): Binomial model prior with model size = 4; Beta-binomial: Beta-binomial model prior

Our overall conclusion on the basis of these analyses is that our forecasting model of state NAEP mathematics achievement would achieve optimal long-run predictive performance for both boys and girls under the fixed-prior setting using a binomial model prior with a model size of four regardless of the choice of the *g*-prior. Clearly, these conclusions rest on our choice of functional form for growth and our choice of predictors used in this analysis.

Summary and conclusions

In this paper, we provided a workflow that permits Bayesian probabilistic forecasting to be applied to large-scale assessments, with NAEP being used as an example. In particular, we borrowed from techniques derived from economic forecasting to specify a forecast model of mathematics achievement across the states and the District of Columbia. Our results demonstrated that a great deal can be learned from applying Bayesian methods to large-scale assessment trend data and that, in particular, the Bayesian perspective provides a rich description of growth while at the same time addressing subjective uncertainty in growth parameters as well as predictive models of growth.

A critical assumption

Throughout this paper, a subtle assumption was invoked but not discussed; namely that the true model for growth, say, M_T was one of the models in the set of models $\{M_k\}_{k=1}^{K}$ that were averaged. This assumption is referred to as the $\mathcal{M}$-*closed* framework, discussed in Bernardo and Smith (2000) and Clyde and Iversen (2013) (see also Kaplan, 2021). Under the $\mathcal{M}$-*closed* framework it makes sense to assign prior probabilities to the space of models reflecting ones belief that the true model M_T is in the space of models under consideration. In fact, this is the framework that underlies the standard approach to BMA discussed in this paper; prior probabilities are assigned to the set of models (typical the indifference prior $1/M$, but others could be chosen, e.g. Fernández et al. (2001a) encoding ones belief that each model is equally likely to be the true model. The application of the indifference prior is the conventional default in Bayesian model averaging software.

In principle, the $\mathcal{M}$-*closed* framework is difficult to warrant and may only really be sensible in the case of simulation studies. Nevertheless, as pointed out by Bernardo and Smith (2000), it may be reasonable to act as though there is a true model. A situation in which one might feel comfortable acting as if a true model exists is when a model has demonstrated good predictive capabilities under a wide variety of situations, but that under a new situation, new uncertainties arise. As this paper is a demonstration of probabilistic forecasting with an application to NAEP, prior experience with this model is not available, and so operating as if $\mathcal{M}$-closed holds is not reasonable.

In contrast to the $\mathcal{M}$-*closed* framework, two other frameworks can be adopted that have important consequences for the predictive modeling described in this paper and represent directions for future research. Both alternatives require the analyst to decide whether to formulate an actual belief model or not. In the first instance the analyst does have an actual belief model but that the models under consideration are proxies for the actual belief model. This is referred to as $\mathcal{M}$-*completed* (Bernardo and Smith, 2000).

Under $\mathcal{M}$-*completed*, the models in $\{M_k\}_{k=1}^{K}$ are listed out for comparison purposes in light of an actual belief model.

In the second instance, the analyst does not even entertain the existence of a belief model; the models in $\{M_k\}_{k=1}^{K}$ are enumerated for comparison purposes only. This is referred to as $\mathcal{M}$-*open* framework. The $\mathcal{M}$-*open* framework is, arguably, the most realistic situation for the social and behavioral sciences. An example of a situation in which $\mathcal{M}$-*open* represents a realistic modeling framework is in specifying a regression model with different choices of predictors. When utilizing BMA to search through the space of possible models (different choices of predictors), it does not make sense to assign priors to the space of models, when there is no actual belief model (or for that matter, a true model) being contemplated.

The distinction among these modeling frameworks is quite important, and indeed, work by Clyde and Iversen (2013) have used a decision-theoretic framework that allows for multi-model inference within the $\mathcal{M}$-*open* framework. This is the so-called method of *stacking* which can be considered either from the frequentist or Bayesian perspective (see e.g. Breiman, 1996; Le and Clarke, 2017; Wolpert, 1992; Yao et al., 2018). We have proceeded under the $\mathcal{M}$-*closed* framework recognizing that this may be unrealistic. The hope is that as predictive modeling of growth using NAEP or other LSA data proceeds, that the performance of these models will lead to a better assessment of whether a true model is at least worth entertaining or whether better forecast performance can be obtained using methods developed for the $\mathcal{M}$-*open* framework.

Frequentist approaches

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.

Policy implications and research directions

It is important to emphasize that the results of this particular study must be treated with great caution because the variables assembled to provide predictors of growth in 8th grade mathematics achievement were few in number and not conceived or designed to provide policy-relevant predictors of growth over time. Going forward, if there is policy interest in using state NAEP data or other large-scale assessments such as PISA or TIMSS for developing predictive models of growth in academic outcomes, then it will become necessary to consider the development of policy-relevant indicators specifically of growth. Nevertheless, the policy implications of the present work lie in the potential for constructing forecast models to guide education policy. For example, for this case

study we find that the changes in reading achievement (as measure here by the difference between 2017 and 2003 reading scores) is the strongest predictor of the growth in math achievement for boys and girls, accounting for parameter and model uncertainty and the specific setting of priors for the forecast models. Such a finding could allow policymakers to forecast changes in math outcomes under different scenarios representing anticipated changes in reading performance over time. Such forecast models can be developed in an ex-post fashion, as shown in this paper, where models are specified and evaluated on the extant data, but eventually utilized in an ex-ante fashion in which a true forecast is generated. Comparison of the forecast to the actual outcome would aid in the iterative calibration of the model (see Dawid, 1982; Little, 2011, for a discussion on calibrated Bayes). It may also be necessary to track the long-run predictive performance of a number of forecasting models under different initial conditions. In any event, a well-calibrated forecasting model could help provide policymakers with additional information needed to make decisions regarding the allocation of funds for interventions or remedial educational programs.

We can also envision future research examining how the predictions arising from different forecasting models can be *ensembled* to provide even better forecasting performance. Ensemble forecasting methods have been developed in contexts such as weather forecasting (e.g. Gneiting and Raftery, 2005), and our current research is being directed to examining the ensemble method of *stacking* (see e.g. Yao et al., 2018) as a means of combining Bayesian predictive distributions in the $\mathcal{M}$-*open* framework with applications to forecasting models using NAEP and other large-scale assessments. An example of stacking applied to PISA 2018 can be found in Kaplan (2021). The end result of research and development into forecasting models for LSAs would be to build one or more usable models that education policy-makers and other stakeholders can use to not just track changes in educational outcomes over time, but to predict the direction and rate of those changes.

To conclude, we argue that large-scale assessments in general, and NAEP in particular are not being sufficiently leveraged for the purposes for which they were created—namely, monitoring population-level trends in achievement. The richness of the trend data that NAEP and other LSAs provide can be utilized for much more than simple, albeit informative, descriptive plots (Kaplan and Jude, in press). The methodological advancements presented in this paper were designed to demonstrate the richness of policy information that can be obtained when using Bayesian predictive models to study educational trends.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-021-00108-2.

Additional file 1. Predictive densities for US States and Jurisdictions.

Acknowledgements
The authors would like to thank Mary Ann Fox, Jiayi Li, Mary Smith, Landa Spingler, Emmanual Sikali and Sinan Yavuz for valuable assistance on this project.

Authors' contributions
DK conceptualized the study and guided the design of the study, the statistical analysis, and contributed to drafting the manuscript. MH carried out the analysis as well as contributed to drafting the manuscript. Both authors read and approved the final manuscript.

Funding
An earlier version of this paper was prepared as a report for the National Center for Education Statistics under Contract No. ED-IES-12-D-0002 with American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Availability of software and data code
The datasets and software supporting the conclusions of this article are available at http://bmer.wceruw.org/index.html.

Declarations

Consent to participate
Not applicable.

Competing interests
The authors declare that there are no competing interests.

Received: 10 December 2020 Accepted: 9 July 2021
Published online: 19 July 2021

References

Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson & S. E. Feinberg (Eds.), *A celebration of statistics* (pp. 1–24). New York: Springer.
Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332.
Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418.
Bernardo, J., & Smith, A. F. M. (2000). *Bayesian theory*. New York: Wiley.
Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.
Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring of education systems: International comparisons. *The ANNALS of the American Academy of Political and Social Science*. https://doi.org/10.1177/0002716219843804.
Breiman, L. (1996). Stacked regressions. *Machine Learning, 24,* 49–64.
Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7,* 434–455.
Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.
Clyde, M. (1999). Bayesian model averaging and model search strategies. In *Bayesian statistics 6* (p. 157–185). Oxford: Oxford University Press.
Clyde, M., & Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. In *Bayesian theory and applications* (p. 483–498). Oxford: Oxford University Press.
Common Core of Data. (2020). Retrieved 20 April, 2020 from https://nces.ed.gov/ccd/.
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,* 278–202.
Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B), 57,* 55–98.
Draper, D., 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. (N-2683-RC). Retrieved from https://www.rand.org/pubs/notes/N2683.html.
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. (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics, 100,* 381–427.
Fernández, C., Ley, E., & Steel, M. F. J. (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics, 16,* 563–576.
Fernández, C., Ley, E., & Steele, M. F. J. (2001c). 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.
Gelman, A., Carlin, J. B., Stern, D. B., Dunson, H. S., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). London: Chapman & Hall.
Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2,* 1360–1383.
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–511.
Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can generally only be understood in the context of the likelihood. *Entropy*. https://doi.org/10.3390/e19100555.
George, E., & Foster, D. (2000). 12, Calibration and empirical Bayes variable selection. *Biometrika, 87,* 731–747. 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. (2005). Weather forecasting with ensemble methods. *Science, 310*(5746), 248–249. https://doi.org/10.1126/science.1115255.
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.
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. Retrieved from http://www.jstor.org/stable/2985032.
Hansen, M., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96,* 746–774. https://doi.org/10.1198/016214501753168398.
Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98,* 879–899.
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14,* 382–417.
Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court.
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. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Newbury Park: Sage Publications.
Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York: Guilford Press.
Kaplan, D. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika*. https://doi.org/10.1007/s11336-021-09754-5.
Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research, 49,* 505–517.
Kaplan, D., & George, R. (1998). Evaluating latent growth models through ex post simulation. *Journal of Educational and Behavioral Statistics, 23,* 216–235.
Kaplan, D., & Jude, N. (in press). Trend analysis with international large-scale assessments: Past practice, current issues, and future directions. In T. Nilsen, A. Stancel-Piątak & J. E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods, findings*. Heidelberg: 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., & Lee, C. (2018). Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review*. 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*. 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.
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.
Laplace, P. S. (1774/1951). *Essai philosophique sur les probabilities*. New York: Dover.
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.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410–423.
Little, R. J. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science, 26,* 162–174.
Madigan, D., Raftery, A. E., & OECD. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical Association, 89,* 1535–1546.
Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04.
Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis, 10,* 292–304.
Mullis, I. V. S. (2013). Foward. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks*. Boston, MA, TIMSS & PIRLS International Study Center Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: Organization for Economic Cooperation and Development.
Plummer, M. (2014). *rjags: Bayesian graphical models using mcmc [Computer software manual]* (R package version 3-13). Retrieved from http://CRAN.R-project.org/package=rjags.
Raftery, A. E. (1998). *Bayes factors and the BIC: Comment on Weakliem* (Tech. Rep. No. 347). University of Washington, Department of Statistics.
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.
Raudenbush, S. W., Bryk, A. S., & OECD. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousands Oaks: Sage Publications.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review, 141*, 2107–2119.
Statisticat, & LLC,. (2020). *Laplacesdemon: Complete environment for bayesian inference [Computer software manual]* (R package version 16.1.4). Retrieved from https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software.
Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature, 58*, 644–719.
UN General Assembly. (2015). *UN General Assembly, Transforming our world: The 2030 Agenda for Sustainable Development, 21 October 2015, A/RES/70/1*. Retrieved 15 June, 2019 from https://www.refworld.org/docid/57b6e3e44.html.
US Department of Education. (2019). *NAEP: Nations Report Card*. Retrieved 16 November, 2019 from https://nces.ed.gov/nationsreportcard/.
US Department of Education. (2021). *NAEP Data Explorer*. Retrieved 6 April, 2021 from https://nces.ed.gov/nationsreportcard/data/.
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., & OECD. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection, and classification tool for microarray data. *Bioinformatics, 21*, 2394–2402.
Zellner, A., & OECD. (1986). On assessing prior distributions and Bayesian regression analysis with *g* prior distributions. Studies in Bayesian econometrics. In P. Goel, A. Zellner & OECD (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233–243). Elsevier: New York.
Zeugner, S., Feldkircher, M., & OECD. (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.

Publisher's Note