Kaplan *et al.*
Large-scale Assessments in Education (2023) 11:2
https://doi.org/10.1186/s40536-022-00140-w

Large-scale Assessments in Education

METHODOLOGY Open Access

Bayesian historical borrowing with longitudinal large-scale assessments

David Kaplan[1*], Jianshen Chen[2], Weicong Lyu[1] and Sinan Yavuz[1]

*Correspondence:
david.kaplan@wisc.edu

[1] Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI, USA
[2] Psychometrics, Learning and Assessment, College Board, New York, NY, USA

Abstract

The purpose of this paper is to extend and evaluate methods of *Bayesian historical borrowing* applied to longitudinal data with a focus on parameter recovery and predictive performance. Bayesian historical borrowing allows researchers to utilize information from previous data sources and to adjust the extent of borrowing based on the similarity of current data to historical data. We examine the utility of three *static* historical borrowing methods including complete pooling, Bayesian synthesis with aggregated data-dependent priors, traditional power priors, and two *dynamic* borrowing methods including Bayesian dynamic borrowing and commensurate priors. Using data from two administrations of the United States Early Childhood Longitudinal Study, we evaluate these methods in terms of in-sample simulation statistics, as well as pseudo out-of-sample measures of predictive performance. A case study examining growth in reading competency over time revealed that for one historical cycle, most methods of historical borrowing perform similarly with respect to predictive performance and parameter recovery. Pooling and power priors performed relatively poorly across the conditions in this study, particularly when the current data and the historical data were heterogeneous. Results from a comprehensive simulation study revealed that the advantages of different historical borrowing methods vary across different evaluation criteria. Overall, Bayesian dynamic borrowing and commensurate priors are no worse, and in some cases better, than other methods in terms of parameter recovery and predictive performance, and considering a previous paper by Kaplan et al. (Psychometrika, 10.1007/s11336-022-09869-3, 2022) found clear benefits of Bayesian dynamic borrowing over other methods of historical borrowing in the multilevel context using data from the Program for International Student Assessment (PISA) with five historical cycles, this paper argues that Bayesian dynamic borrowing or commensurate priors is a prudent choice for borrowing information from previous cycles of longitudinal data.

Introduction

With the increased availability of large-scale longitudinal surveys and assessments, researchers can address critical questions of growth and development. Such endeavors as the German National Educational Panel Study (NEPS) (Blossfeld & Roßbach, 2019) or the US Early Childhood Longitudinal Studies Program (ECLS; NCES, 2018) provide extensive data on the academic, behavioral, and social development of children throughout the school years and beyond. These data can be used to estimate rates of change in

Springer Open

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

relevant academic and non-academic outcomes over the waves of the study as well as to model variation in rates of change over children as a function of a very large number of covariates collected on children, their families, and their schools.

In addition to providing estimated rates of growth in academic and non-academic outcomes, longitudinal studies can be leveraged for the purposes of forecasting growth beyond the extant waves of the study. Although such uses are rare in education (see however Kaplan & Huang, 2021), it still remains an important use of longitudinal data. This paper is concerned with the question of whether it is possible and desirable to borrow information from an analysis of the waves of a previous cycle of longitudinal data to inform the analysis of the waves of a current longitudinal study, particularly with respect to the estimation of growth and predictive performance. With a focus on the ECLS Kindergarten cohorts (ECLS-K), the question concerns specifically whether estimation of the growth rates in reading performance from the 2010 to 2011 cohort (referred to as ECLS-K:2010-11) can be improved by systematically including information about the growth rates from the 1998 to 1999 cohort (referred to as ECLS-K:1998-99).

The methods we examine in this paper are generally referred to as *Bayesian historical borrowing*, a class of procedures that have long been applied in the clinical trials field (e.g. Pocock, 1976; Hobbs et al., 2011; Hobbs et al., 2012; Schmidli et al., 2014; Viele et al., 2014), and recently developed and applied to large-scale and cross-sectional educational data (see Kaplan et al., 2022). Two general approaches to historical borrowing can be identified in the literature. The first are *static borrowing* methods where prior strength does not automatically vary based on the similarity between the historical data and the current data. For example, with static borrowing, fixed prior strength might be based on a researcher's judgement regarding the similarity between the historical data and current data, but this prior strength would not automatically be adjusted based on the heterogeneity between historical data and current data to supplement the researcher's judgement. Methods under static borrowing that we will study in this paper include *pooling*, also known as *integrative data analysis* (Bainter & Curran, 2015; Curran & Hussong, 2009), *Bayesian synthesis* using augmented or aggregated data-dependent priors (Marcoulides, 2017), and *power priors* (Ibrahim & Chen, 2000; Chen et al., 2000; Chen et al., 2015).

In contrast to static borrowing, *dynamic borrowing* methods allow for a joint prior distribution to be specified over both the historical and current data to encode the researcher's judgement regarding the similarity between the historical data and the current data. The similarity is controlled by the variance of the joint prior distribution. Methods under dynamic borrowing that will be examined in this paper include *Bayesian dynamic borrowing* (see Viele et al., 2014; Kaplan et al., 2022) and *commensurate priors* (Hobbs et al., 2011, 2012). For a review of the static and dynamic borrowing methods examined in this paper, one may refer to Kaplan et al. (2022).

The organization of this paper is as follows. In the next section, we specify growth curve modeling as a Bayesian hierarchical model. Although growth curve modeling is not the only method that can be applied to longitudinal data, we focus on this method because of its extensive use and its flexibility in addressing specific issues of growth. In the following two sections, we extend static and dynamic borrowing methods to growth curve models. Following this, we then discuss the methodology for how to evaluate such

models - particularly using statistics that assess in-sample and pseudo out-of-sample predictive performance. This is then followed by the design and results of our case study and simulation study. Future research directions are outlined in the Conclusions section.

Bayesian growth curve modeling

To fix terminology and notation, we use the term *cycle* to refer to the different survey administrations (i.e. ECLS-K:1998-99 and ECLS-K:2010-11), and use the term *wave* to refer to the repeated measurements within the cycle. Let superscript 0 represent the current cycle, ECLS-K:2010-11, and h ($h = 1 \ldots H$) represent the historical cycle(s), where for our paper $H = 1$ with ECLS-K:1998-99 as the only historical cycle. The level-1 (within-student) model can be written as follows. Let

$$\boldsymbol{y}_{ig}^{0} = \boldsymbol{\Lambda}^{0}\, \boldsymbol{\eta}_{ig}^{0} + \boldsymbol{\epsilon}_{y_{ig}}^{0} \tag{1}$$

where $\boldsymbol{y}_{ig}^{0}$ be a $T \times 1$ vector of T waves of measurement for student i ($i = 1, \ldots, n_g$) in school g ($g = 1, \ldots G$); $\boldsymbol{\Lambda}^{0}$ be a $T \times Q$ matrix of fixed constants that, for notational simplicity, are assumed to be the same for all participants across all schools. These fixed constants serve to parameterize the growth model as a structural equation model (see Willett & Sayer, 1994). Further, let $\boldsymbol{\eta}_{ig}^{0}$ be a $Q \times 1$ vector of random growth parameters for student i in school g. In our study, $Q = 3$, representing the intercept, linear growth component, and quadratic growth component for each student in each school; and $\boldsymbol{\epsilon}_{y_{ig}}^{0}$ is a $T \times 1$ vector of residuals, with diagonal covariance matrix $\boldsymbol{\Sigma}_{y}^{0}$. Here, we assume constant variances across students and schools, and so

$$\boldsymbol{\Sigma}_{y}^{0} = \begin{bmatrix} \sigma_{y_1}^2 & & & & \\ & \sigma_{y_2}^2 & & & \\ & & \ddots & & \\ & & & \sigma_{y_{T-1}}^2 & \\ & & & & \sigma_{y_T}^2 \end{bmatrix}. \tag{2}$$

The level-2 (between-student) model allows the random growth parameters to be related to a set of time-invariant predictors. The level-2 model can be written as

$$\boldsymbol{\eta}_{ig}^{0} = \boldsymbol{\Gamma}_{g}^{0} \boldsymbol{x}_{ig}^{0} + \boldsymbol{\epsilon}_{\eta_{ig}}^{0}, \tag{3}$$

where $\boldsymbol{\Gamma}_{g}^{0}$ is a $Q \times P$ matrix of regression coefficients associated with the time-invariant predictors which vary over schools, $\boldsymbol{x}_{ig}^{0}$ is a $P \times 1$ vector of time-invariant predictors whose values vary over students and schools, and $\boldsymbol{\epsilon}_{\eta_{ig}}^{0}$ is a $Q \times 1$ (i.e., intercept, slope and quadratic growth component in our study with $Q = 3$) vector of residuals with symmetric covariance matrix $\boldsymbol{\Sigma}_{\eta}^{0}$. We assume constant variances across students and schools, and so

$$\boldsymbol{\Sigma}_{\eta}^{0} = \begin{bmatrix} \sigma_{\eta_0^0}^2 & \sigma_{\eta_1^0 \eta_0^0} & \sigma_{\eta_2^0 \eta_0^0} \\ \sigma_{\eta_1^0 \eta_0^0} & \sigma_{\eta_1^0}^2 & \sigma_{\eta_2^0 \eta_1^0} \\ \sigma_{\eta_2^0 \eta_0^0} & \sigma_{\eta_2^0 \eta_1^0} & \sigma_{\eta_2^0}^2 \end{bmatrix}, \tag{4}$$

which allows for the random growth parameters (i.e., intercept, slope, and quadratic components) to be correlated conditional on the predictors as shown in Eq. (4). The assumption of constant $\boldsymbol{\Sigma}_{\eta}^{0}$ could, in principle, be relaxed.

The level-3 (between-school) model can be written as

$$\text{vec}(\boldsymbol{\Gamma}_{g}^{0}) = \boldsymbol{\Pi}^{0}\,\boldsymbol{w}_{g} + \boldsymbol{\epsilon}_{\Gamma_{g}}^{0} \tag{5}$$

where $\text{vec}(\cdot)$ is the vectorization operator that turns a matrix into a long column vector, $\boldsymbol{\Pi}^{0}$ is a $QP \times M$ matrix of between-school regression coefficients, $\boldsymbol{w}_{g}$ is a $M \times 1$ vector of school-level predictors, and $\boldsymbol{\epsilon}_{\Gamma_{g}}^{0}$ is a $QP \times 1$ vector of residuals with covariance matrix $\boldsymbol{\Sigma}_{\Gamma}^{0}$. We assume constant variances across schools, and so

$$\boldsymbol{\Sigma}_{\Gamma}^{0} = \begin{bmatrix} \sigma^2_{\Gamma_1} & & & & \\ & \sigma^2_{\Gamma_2} & & & \\ & & \ddots & & \\ & & & \sigma^2_{\Gamma_{QP-1}} & \\ & & & & \sigma^2_{\Gamma_{QP}} \end{bmatrix}. \tag{6}$$

Bayesian hierarchical growth curve model

A Bayesian hierarchical specification of the growth curve model in Eqs. (1–6) can be written as

$$\boldsymbol{y}_{ig}^{0} \sim N(\boldsymbol{\Lambda}^{0}\boldsymbol{\eta}_{ig}^{0}, \boldsymbol{\Sigma}_{y}^{0}), \tag{7a}$$

$$\boldsymbol{\eta}_{ig}^{0} \sim N(\boldsymbol{\Gamma}_{g}^{0}\boldsymbol{x}_{ig}^{0}, \boldsymbol{\Sigma}_{\eta}^{0}), \tag{7b}$$

$$\text{vec}(\boldsymbol{\Gamma}_{g}^{0}) \sim N(\boldsymbol{\Pi}^{0}\,\boldsymbol{w}_{g}^{0}, \boldsymbol{\Sigma}_{\Gamma}^{0}), \tag{7c}$$

$$\text{vec}(\boldsymbol{\Pi}^{0}) \sim N(\boldsymbol{\Omega}^{0}, \boldsymbol{\Sigma}_{\Pi}^{0}) \tag{7d}$$

where $\boldsymbol{\Omega}^{0}$ and $\boldsymbol{\Sigma}_{\Pi}^{0}$ are fixed and known parameters. Prior distributions on the residual covariance matrices are assumed to be inverse-Wishart (IW). For the current data, following the notation in Eq. (2), we assume all the diagonal elements of $\boldsymbol{\Sigma}_{y}^{0}$ are equal to $(\sigma_{y}^{0})^2$, where σ_{y}^{0} indicates the standard deviation of level-1 residual. The prior distributions for the variation parameters are specified as follows:

$$\sigma_{y}^{0} \sim \text{ half-Cauchy } (0, 1), \tag{8a}$$

$$\boldsymbol{\Sigma}_{\eta}^{0} \sim \text{ IW } (\mathbf{R}_{\eta}, \nu_{\eta}), \tag{8b}$$

$$\boldsymbol{\Sigma}_{\Gamma}^{0} \sim \text{ IW } (\mathbf{R}_{\Gamma}, \nu_{\Gamma}). \tag{8c}$$

Static borrowing for growth curve models

Extensions of growth curve models to pooling and to aggregated data-dependent priors are relatively straightforward. In particular, assuming the conditions for pooling of longitudinal data are met (see Hofer & Piccinin, 2009), pooling of data would be relatively straightforward. Similarly, the aggregated data-dependent prior approach of Marcoulides (2017) would require obtaining averages of the parameters of interest from the historical cycles and implementing them as the hyperparameters of informative prior distributions for the current cycle. This section concentrates instead on expanding the power prior to longitudinal data.

The power prior for growth curve models

Consider again the Bayesian hierarchical specification of the GCM models in Eqs. (7a–7d). Among the entire set of model parameters, the top level parameters in the growth curve models $\mathbf{\Pi}$ are the common parameters across cycles that will borrow from historical data through power priors. The lower level parameters are cycle specific and thus there is no direct borrowing. The probability distribution of the historical data given both the common and the unique parameters across cycles can be written as

$$p(\boldsymbol{y} \mid \boldsymbol{\eta}, \mathbf{\Gamma}, \mathbf{\Sigma}_y, \mathbf{\Sigma}_\eta, \mathbf{\Sigma}_\Gamma, \mathbf{\Pi}) = \prod_{h=1}^{H} \prod_{g=1}^{G^h} \prod_{i=1}^{n_g^h} p(y_{ig}^h \mid \boldsymbol{\eta}_{ig}^h, \mathbf{\Sigma}_y^h) p(\boldsymbol{\eta}_{ig}^h \mid \mathbf{\Gamma}_g^h, \mathbf{\Sigma}_\eta^h) p(\mathbf{\Gamma}_g^h \mid \mathbf{\Sigma}_\Gamma^h, \mathbf{\Pi}). \tag{9}$$

From here, the power prior can be expressed as

$$p(\mathbf{\Pi} \mid \boldsymbol{y}^1, \ldots, \boldsymbol{y}^h, a) \propto \left[\prod_{h=1}^{H} \prod_{g=1}^{G^h} \prod_{i=1}^{n_g^h} p(y_{ig}^h \mid \boldsymbol{\eta}_{ig}^h, \mathbf{\Sigma}_y^h) p(\boldsymbol{\eta}_{ig}^h \mid \mathbf{\Gamma}_g^h, \mathbf{\Sigma}_\eta^h) p(\mathbf{\Gamma}_g^h \mid \mathbf{\Sigma}_\Gamma^h, \mathbf{\Pi}) \right]^a p(\mathbf{\Pi}), \tag{10}$$

where a controls the strength of borrowing from historical data on $p(\mathbf{\Pi} \mid \boldsymbol{y}^1, \ldots, \boldsymbol{y}^h, a)$. Notice that when $a = 0$, the prior does not depend on the historical data, and when $a = 1$, the prior is the posterior distribution from the previous study.

Dynamic borrowing for growth curve models

As noted earlier, static borrowing methods do not incorporate information about the current cycle into the prior specification. In contrast, dynamic borrowing methods do incorporate the current cycle into the prior specification of the model parameters. In this section, we extend methods of dynamic borrowing to growth curve models, concentrating on Bayesian dynamic borrowing and commensurate priors.

Extensions of Bayesian dynamic borrowing to growth curve models

We adapt the cross-sectional multilevel modeling notation of Kaplan et al. (2022) to the case of growth curve models. We begin by borrowing from historical cycles to estimate the growth parameters. This requires defining a joint distribution of the growth parameters over the historical cycles (denoted as cycles 1 to H) and the current cycle (denoted as cycle 0), which is assumed to be a multivariate Guassian distribution with

the $Q(H+1) \times 1$ mean vector $\boldsymbol{\mu}_{\eta_{ig}}$ and $Q(H+1) \times Q(H+1)$ block-diagonal covariance matrix $\mathbf{T}_{\eta}$.

$$\text{vec}\left(\boldsymbol{\eta}_{ig}^{0}, \boldsymbol{\eta}_{ig}^{1}, \ldots, \boldsymbol{\eta}_{ig}^{H-1}, \boldsymbol{\eta}_{ig}^{H}\right) \sim N(\boldsymbol{\mu}_{\eta_{ig}}, \mathbf{T}_{\eta}), \tag{11}$$

where following Eq. (7b), $\boldsymbol{\mu}_{\eta_{ig}} = \text{vec}\left(\boldsymbol{\Gamma}_{g}^{0}\boldsymbol{x}_{ig}^{0}, \boldsymbol{\Gamma}_{g}^{1}\boldsymbol{x}_{ig}^{1}, \ldots, \boldsymbol{\Gamma}_{g}^{H}\boldsymbol{x}_{ig}^{H}\right)$, $\boldsymbol{\Gamma}_{g}^{h}$ $(h = 1, 2, \ldots, H)$ represents a $Q \times P$ vector of the time-invariant regression coefficients for the h^{th} historical cycle and $\boldsymbol{\Gamma}_{g}^{0}$ represents a $Q \times P$ vector of the current time-invariant regression coefficients.

The covariance matrix of the random growth parameters can be written as

$$\mathbf{T}_{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}_{\eta}^{0} & & & & \\ & \boldsymbol{\Sigma}_{\eta}^{1} & & & \\ & & \ddots & & \\ & & & \boldsymbol{\Sigma}_{\eta}^{H-1} & \\ & & & & \boldsymbol{\Sigma}_{\eta}^{H} \end{bmatrix}, \tag{12}$$

where each element of Eq. (12) is a symmetric matrix as in Eq. (4).

Next, the joint distribution of $\boldsymbol{\Gamma}_{g}^{0}, \boldsymbol{\Gamma}_{g}^{1}, \ldots, \boldsymbol{\Gamma}_{g}^{H}$ are assumed to be multivariate Gaussian with the $QP(H+1) \times 1$ mean vector $\boldsymbol{\mu}_{\Gamma_g}$ and $QP(H+1) \times QP(H+1)$ block-diagonal covariance matrix $\mathbf{T}_{\Gamma}$ —viz.

$$\text{vec}(\boldsymbol{\Gamma}_{g}^{0}, \boldsymbol{\Gamma}_{g}^{1}, \ldots, \boldsymbol{\Gamma}_{g}^{H}) \sim N(\boldsymbol{\mu}_{\Gamma_g}, \mathbf{T}_{\Gamma}), \tag{13}$$

where following Eq. (7c), $\boldsymbol{\mu}_{\Gamma_g} = \text{vec}\left(\boldsymbol{\Pi}^{0}\boldsymbol{w}_{g}^{0}, \boldsymbol{\Pi}^{1}\boldsymbol{w}_{g}^{1}, \ldots, \boldsymbol{\Pi}^{H}\boldsymbol{w}_{g}^{H}\right)$, $\boldsymbol{\Pi}^{h}$ $(h = 1, \ldots H)$ represents a $QP \times M$ matrix of the school-level regression coefficients for the h^{th} historical cycle, and $\boldsymbol{\Pi}^{0}$ represents a $QP \times M$ matrix of the school-level regression coefficient for the current cycle.

The covariance matrix of the time-invariant regression coefficients over the current and historical cycles, $\mathbf{T}_{\Gamma}$, is specified as being block diagonal,

$$\mathbf{T}_{\Gamma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\Gamma}^{0} & & & & \\ & \boldsymbol{\Sigma}_{\Gamma}^{1} & & & \\ & & \ddots & & \\ & & & \boldsymbol{\Sigma}_{\Gamma}^{H-1} & \\ & & & & \boldsymbol{\Sigma}_{\Gamma}^{H} \end{bmatrix}, \tag{14}$$

where the elements of $\mathbf{T}_{\Gamma}$ contain the variances and covariances of the regression coefficients within each historical or current data set. We assume that elements outside the block diagonal of $\mathbf{T}_{\Gamma}$ are null matrices.

Finally, the joint distribution of $\boldsymbol{\Pi}^{0}, \boldsymbol{\Pi}^{1}, \ldots, \boldsymbol{\Pi}^{H}$ is also assumed to be multivariate Gaussian with a $QPM \times 1$ mean vector $\boldsymbol{\mu}_{\Pi}$ and $QPM \times QPM$ covariance matrix $\mathbf{T}_{\Pi}$ —namely

$$\text{vec}\left(\boldsymbol{\Pi}^{0}, \boldsymbol{\Pi}^{1}, \ldots, \boldsymbol{\Pi}^{H}\right) \sim N(\boldsymbol{\mu}_{\Pi}, \mathbf{T}_{\Pi}). \tag{15}$$

The covariance matrix $\mathbf{T}_\Pi$ can be diagonal with elements τ^2, which controls the degree of borrowing across cycles. Note that $\mathbf{\Pi}^0, \mathbf{\Pi}^1, \ldots, \mathbf{\Pi}^H$ follows the same mean vector $\boldsymbol{\mu}_\Pi$ and covariance matrix $\mathbf{T}_\Pi$ as shown in Eq. (15) and thus the elements of $\boldsymbol{\mu}_\Pi$ and $\mathbf{T}_\Pi$ are not cycle specific, indicating that borrowing across cycles takes place at the top level of the hierarchy.

Commensurate priors

Hobbs et al. (2011) proposed dynamic versions of power priors, referred to as *commensurate power priors*, where the coefficient used to downweight the historical data is viewed as random and estimated based on a measure of the agreement between the current and historical data. Hobbs et al. (2011) also proposed general commensurate priors where the prior mean for the current parameters of interest is conditioned on the historical population mean and the prior precision τ, referred to as the *commensurability* parameter, which reflects the commensurability between the current and historical parameters.[1] Hobbs et al. (2011) evaluated commensurate priors in a scenario of borrowing one historical trial to analyze a single-arm trial, that is, assuming there is only one historical study and one parameter of interest, β. The location parameter or mean for β is μ_β^0 for the current data and μ_β^H for the historical data. Then the commensurate prior for μ_β^0 can be specified as $\mu_\beta^0 \sim N(\mu_\beta^H, 1/\tau)$.

As discussed in Hobbs et al. (2012), the commensurate prior in Hobbs et al. (2011) suffers from the fact that diffuse priors could actually become undesirably informative and that the historical likelihood is considered as a component of the prior rather than data. Therefore, Hobbs et al. (2012) proposed a modified commensurate prior that incorporates historical data as part of the likelihood for the current parameter estimation and employs empirical and fully Bayesian modifications for estimating the commensurate parameter τ (e.g., as illustrated in Eq. 1 of their paper). They also extended the method to general and generalized linear mixed regression models in the context of two successive clinical trials.

The modified commensurate priors approach in Hobbs et al. (2012) was compared to several meta-analytic models where priors for the historical parameters and current parameters were jointly modeled, but historical data were not incorporated in the likelihood of the current parameter estimation and thus the priors were not commensurate or dynamic. Commensurate priors were shown to provide more bias reduction compared to the meta-analytic approaches they evaluated. The bias reduction was larger when there was only one historical study compared to when there were two or three historical studies.

Although Kaplan et al. (2022) extended Bayesian dynamic borrowing to cross-sectional single-level and multilevel models with covariates, they did not examine commensurate priors. For this paper, we consider the modified commensurate prior in Hobbs et al. (2012) and implement it in the multilevel setting with multiple historical studies in the following way. For regression coefficients, we assume:

[1] Note that in Hobbs et al. (2011; 2012), τ refers to prior precision. But in this paper, we use τ^2 to indicate prior variance following the notation in Viele et al. (2014).

$$\boldsymbol{\beta}^1 = \cdots = \boldsymbol{\beta}^H = \boldsymbol{\beta}^{Hist}, \tag{16}$$

where $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^H$ are regression coefficients of interest for each historical cycle. Although regression coefficients are likely to be different within a historical cycle, the common regression coefficients (e.g., intercepts) are assumed to be equal across historical cycles and denoted as $\boldsymbol{\beta}^{Hist}$, where $\boldsymbol{\beta}^{Hist}$ can be given a vague Gaussian prior.

The parameters for the current cycle (denoted as cycle 0) follow a prior distribution with the historical regression coefficients as the prior mean as follows:

$$\boldsymbol{\beta}^0 \sim N(\boldsymbol{\beta}^{Hist}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{17}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ can be specified as a diagonal matrix, diag $(\sigma_1^2, \ldots, \sigma_P^2,)$, for P regression coefficients, and each element of the diagonal matrix can be provided its own prior distribution, such as inverse-gamma (IG), half-Cauchy (see e.g. Gelman, 2006), spike-and-slab (Mitchell & Beauchamp, 1988), etc.

Considering a two-level setting such as students nested in schools, the priors for the school-level covariance matrices for historical cycles can be specified as follows:

$$\boldsymbol{\Sigma}^1 = \cdots = \boldsymbol{\Sigma}^H = \boldsymbol{\Sigma}^{Hist} \tag{18a}$$

$$\boldsymbol{\Sigma}^{Hist} = \boldsymbol{\sigma}\,\boldsymbol{\Omega}\boldsymbol{\sigma} \tag{18b}$$

$$\boldsymbol{\sigma} \sim \text{half-Cauchy}(0, 1) \tag{18c}$$

$$\boldsymbol{\Omega} \sim \text{LKJCorr}(1), \tag{18d}$$

where $\boldsymbol{\Sigma}^1, \ldots, \boldsymbol{\Sigma}^H$ are covariance matrices for each historical cycle. The common elements of the covariance matrices (e.g., variances) are assumed to be equal across historical cycles and denoted as $\boldsymbol{\Sigma}^{Hist}$, and $\boldsymbol{\Omega}$ is a correlation matrix following the Lewandowski et al. 2009 (LKJ) prior.[2]

For the current cycle (denoted as cycle 0), the inverse of covariance matrix can follow a prior distribution conditioning on the historical covariance matrix as

$$(\boldsymbol{\Sigma}^0)^{-1} \sim \text{Wishart}(\nu, \nu(\boldsymbol{\Sigma}^{Hist})^{-1}). \tag{19}$$

For multilevel settings with three levels or more, the higher level covariance matrices may follow the similar prior specifications as the above.

Note that Bayesian dynamic borrowing differs from commensurate priors insofar as the joint prior distribution for the former contains the current cycle, while commensurate priors place a prior on the common historical parameters of interest first and then the current parameter has a prior distribution with the historical regression coefficients as the mean as shown in Eq. (17).

[2] The LKJ correlation prior is suitable as a prior distribution for correlation matrices. Its density satisfies $\text{LKJCorr}(\boldsymbol{\Sigma} \mid \delta) \propto [\det(\boldsymbol{\Sigma})]^{\delta-1}$, so $\delta = 1$ leads to a uniform prior over all possible correlation matrices while $\delta > 1$ leads to a prior that places more mass near the identity matrix (Lewandowski et al., 2009).

Extensions of commensurate priors to growth curve models

Our extension of commensurate priors to growth curve models closely follows the notation for commensurate priors in multilevel settings. We consider a multilevel growth curve model with multiple time points within an individual and individuals nested in groups such as students nested in schools. To simplify the notation, we stack regression coefficients of the growth curve model, including growth parameters and regression coefficients of individual-level time-invariant predictors and group-level predictors, together to be $\boldsymbol{\beta}$. We let $\boldsymbol{\Sigma}_I$ and $\boldsymbol{\Sigma}_G$ denote the corresponding individual-level and group-level covariance matrices.

The prior specification for historical regression coefficients $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^H$ can follow those in Eq. (16) and the prior specification for current regression coefficients $\boldsymbol{\beta}^0$ can follow those in Eq. (17). Similarly, the prior specification for historical covariance matrices $\boldsymbol{\Sigma}_I^1, \ldots, \boldsymbol{\Sigma}_I^H$ and $\boldsymbol{\Sigma}_G^1, \ldots, \boldsymbol{\Sigma}_G^H$ can follow those in Eq. (18) and the prior specification for current covariance matrices $\boldsymbol{\Sigma}_I^0$ and $\boldsymbol{\Sigma}_G^0$ can follow those in Eq. (19).

We introduce a modification to the estimation of the commensurate prior for this study. Instead of using the spike-and-slab prior (Mitchell & Beauchamp, 1988) used by Hobbs et al. (2012) for the commensurability parameter, for computational simplicity and numerical stability, we utilize an extension of the *horseshoe prior* (Carvalho et al., 2010) developed by Piironen and Vehtari (2017) to account for commensurability. The horseshoe prior is a *global-local* shrinkage prior that combines together two priors: a global prior for all of the coefficients in the current cycle, which has the effect of shrinking all coefficients toward historical coefficients, and a local prior for each of the predictors in the current cycle, which has the effect of relaxing the shrinkage due to the global prior for coefficients that are away from historical coefficients.

Following the notation in Betancourt (2018), the horseshoe prior for the p^{th} element of $\boldsymbol{\beta}^0$ can be specified as follows:

$$\beta_p^0 \sim N(0, \tau\lambda_p) \tag{20a}$$

$$\lambda_p \sim \text{ half-Cauchy } (0, 1) \tag{20b}$$

$$\tau \sim \text{ half-Cauchy } (0, \tau_0), \tag{20c}$$

where τ_0 is a hyperparameter that controls the behavior of the global shrinkage prior τ (Carvalho et al., 2010).[3]

A limitation of the conventional horseshoe prior relates to the regularization of the large coefficients. Specifically, it is still the case that large coefficients can transcend the global scale set by τ_0 with the impact being that the posteriors of these large coefficients can become quite diffused, particularly in the case of weakly-identified coefficients (Betancourt, 2018). To remedy this issue, Piironen and Vehtari (2017) proposed a *regularized* version of the horseshoe prior (also known as the *Finnish horseshoe prior*) that has the following form:

[3] The horseshoe prior gets its name from the fact that under certain conditions, the probability distribution of the shrinkage parameter associated with horseshoe prior reduces to a Beta$(\frac{1}{2}, \frac{1}{2})$ distribution, which has the shape of a horseshoe.

$$\beta_p^0 \sim N(0, \tau\tilde{\lambda}_p) \tag{21a}$$

$$\tilde{\lambda}_p = \frac{c\lambda_m}{\sqrt{c^2 + \tau^2\lambda_m^2}} \tag{21b}$$

$$\lambda_p \sim \text{ half-Cauchy } (0, 1) \tag{21c}$$

$$c^2 \sim \text{ inv-gamma } \left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right) \tag{21d}$$

$$\tau \sim \text{ half-Cauchy } (0, \tau_0), \tag{21e}$$

where s^2 is the variance for each of the p predictor variables, assumed to be constant, and c is the slab width. The hyperparameters of the inverse-gamma distribution in Eq. (21d) induce a Student-$t_\nu(0, s^2)$ distribution for the slab (see Piironen & Vehtari, 2017, for more detail). For our paper, two changes were implemented to the regularized horseshoe. First, we set the mean of β_p^0 to β_p^{hist} rather than to zero in order for shrinkage to be toward the historical mean, i.e., $\beta_p^0 \sim N(\beta_p^{hist}, \tau\tilde{\lambda}_p)$. Second, we set $s^2 = 1$ due to standardization of the data.

Evaluating predictions under historical borrowing

For this paper, we evaluate historical borrowing for longitudinal data using two distinct approaches. The first is based on classic approaches developed in the econometrics literature based on so-called *in-sample simulations* wherein the growth record predicted by the model under various methods of borrowing is compared to the actual growth record. This approach was used by Kaplan and George (1998) as a general framework for the evaluation of frequentist growth curve models without relying on conventional goodness-of-fit tests. The second approach is based on the use of scoring rules to evaluate the overall predictive accuracy of probabilistic forecasts. We refer to these methods as *pseudo out-of-sample* performance measures because they will be used to compare the predicted distribution of the outcome from the last wave of the cycle to the known distribution of the outcome.

In-sample simulations

Following closely the discussion in Kaplan and George (1998) in the context of latent variable growth models, an alternative form of model assessment that does not rely solely on conventional goodness-of-fit statistics in structural equation models (see e.g. Kaplan, 2009) concerns how well the model can reproduce the known growth trajectory. In the context of economic forecasting, this type of model evaluation is referred to as *in-sample* simulation (Pindyck & Rubinfeld, 1991). Given the estimated parameters of the model and estimated average values of any exogenous predictors, in-sample simulations can be used to predict the known growth record. These in-sample simulation statistics, to be described below, can then be applied to assess how well the model-based predictions fit the known growth record. The result of such an exercise is a form of model evaluation that goes beyond simply assessing overall goodness-of-fit and considers the utility

of the model for some other purpose. In this case, one might be concerned with simulation adequacy when considering the use of these models for subsequent forecasting.

Based on the early work of Theil (1966, see also; Pindyck & Rubinfeld, 1991; Kaplan & George 1998), we use three in-sample simulation measures to assess how well the fitted growth trajectory compares to the actual growth trajectory. While other measures are available, these three are considered classic measures in the econometric literature.

Theil's inequality coefficient

The first measure we consider is Theil's 1966 inequality coefficient, defined as

$$U = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t^s - y_t^a)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t^s)^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t^a)^2}}, \tag{22}$$

where T is the number of time periods, y^s is the simulated (i.e., predicted) value at time t and y^a is the actual value at time t. In the context of this paper, y^s is the model based predicted value of y at time t under different scenarios of dynamic borrowing, and y^a is the actual mean (over individuals) of y at time t for the current cycle ECLS-K:2010-11. From Eq. (22), it can be seen that a value of $U = 0$ indicates a perfect fit of the simulated growth record to the actual growth record. On the other hand, if $U = 1$, the simulation adequacy is as poor as possible.

Theil's bias proportion

As shown by Theil (1966) and also discussed in Pindyck and Rubinfeld (1991), the inequality coefficient can be decomposed into measures that provide different perspectives on the quality of simulation performance. The first component of Theil's U is the bias proportion, defined as

$$U^B = \frac{(\bar{y}^s - \bar{y}^a)^2}{\frac{1}{T}\sum_{i=1}^{T}(y_t^s - y_t^a)^2}, \tag{23}$$

where $\bar{y}^s$ and $\bar{y}^a$ are the means of the simulated and actual growth record, respectively, calculated across the T time periods. The bias proportion provides a measure of systematic error because it considers deviations of average actual values from average simulated values. The ideal would be a value of $U^B = 0$. A suggested rule of thumb is that values over 0.1 or 0.2 are indicative of systematic bias (Pindyck & Rubinfeld, 1991, p. 341).

Theil's variance proportion

Another component of Theil's U is the variance proportion defined as

$$U^V = \frac{(\sigma^s - \sigma^a)^2}{\frac{1}{T}\sum_{i=1}^{T}(y_t^s - y_t^a)^2}, \tag{24}$$

where σ_s and σ_a are the standard deviations of the simulated and actual growth records, respectively, calculated across the T time periods. The variance proportion provides a measure of the extent to which the model tracks the variability in the growth record. As noted by Pindyck and Rubinfeld (1991), if U^V is large, it suggests that the actual (or simulated) growth record varies a great deal while the simulated (or actual) growth record does not deviate by a comparable amount.[4]

Pseudo out-of-sample scoring rules for probabilistic forecasts

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 prediction can be said to be "well-calibrated" if the assigned probability of the outcome matches the actual proportion of times that the outcome occurred (Dawid, 1982).

A large number of scoring rules have been reviewed in the literature (see e.g., Winkler, 1996; Bernardo & Smith, 2000; Jose et al., 2008; Merkle & Steyvers, 2013; Gneiting & Raftery, 2007). Here, however, we focus on two strictly proper scoring rules that are commonly used to evaluate predictive distributions: the *Kullback-Leibler divergence score* and *expected log point-wise predictive density* implemented through *leave-one-out cross-validation*.

Kullback-Leibler Divergence score

For this paper, we evaluate the quality of predictions using the Kullback-Leibler Divergence (KLD) score (Kullback & Leibler, 1951; Kullback, 1959; 1987). Consider two distributions, $p(y)$ and $g(y|\theta)$, where $p(y)$ could be the distribution of observed reading literacy scores, and $g(y|\theta)$ could be the prediction of these reading scores based on a model. The KLD between these two distributions can be written as

$$\text{KLD}\ (f,g) = \int p(y) \log\left(\frac{p(y)}{g(y|\theta)}\right) dy, \tag{25}$$

where KLD(f, g) is the information lost when $g(y|\theta)$ is used to approximate $p(y)$. For this paper, the actual reading outcome scores will be compared to the predicted outcome using different methods of borrowing from historical data, including not borrowing historical data at all. The model with the lowest KLD measure is deemed best in the sense that the information lost is lowest when approximating the actual reading outcome distribution with the distribution predicted on the basis of the model.

[4] A final measure based on the decomposition of the inequality coefficient is the *covariance proportion* U^C, which measures the error that remains after having removed deviations from average values. According to Theil (1966), it is unreasonable to expect that predicted and actual values will lie on a straight line and thus large values of the covariance proportion are not deemed as serious as large values of U^B of U^V. Note that $U^B + U^V + U^C = 1$. We do not include the covariance proportion measure in this paper.

Leave-one-out cross validation information criterion (LOOIC)

In addition to KLD, we also examine the predictive performance of Bayesian historical borrowing methods using the *leave-one-out cross validation information criterion* (LOOIC). Leave-one-out-cross-validation (LOOCV) is a special case of k-fold cross-validation (k-fold CV) when $k = n$, with n indicating the number of observations. In k-fold CV, a sample is split into k groups (folds) and each fold is taken to be the validation set with the remaining $k - 1$ folds serving as the training set. For LOOCV, each observation serves as the validation set with the remaining $n - 1$ observations serving as the training set. For each observation, the *expected log point-wise predictive density* (ELPD) is calculated and serves as a score of the predictive accuracy for n data points taken one at a time (see Vehtari et al., 2017)[5] An information criterion referred to as the LOOIC can then be obtained as function of the estimated ELPD. Among a set of competing models, the one with the smallest LOOIC is considered the best from an out-of-sample point-wise predictive point of view. For this paper, we evaluated two different LOOIC indices with students being left out one at a time and with schools being left out one at a time. We obtain the Bayesian LOOIC provided by the loo software program (Vehtari et al., 2019), available in R (R Core Team, 2022).

Data source: the Early Childhood Longitudinal Study

This paper will utilize data from the two extant cycles of the Early Childhood Longitudinal Study (ECLS). Specifically, we will focus on the ECLS kindergarten cohort of 2011 (ECLS-K:2010-11) and utilize the ECLS kindergarten cohort of 1998-99 as prior information to inform our simulation studies as well as using these data sources for our case study. The ECLS program was sponsored by the National Center for Education Statistics (NCES) and is a component of the NCES Longitudinal Studies Program. The main purpose of ECLS is to provide policymakers, researchers, and the interested community at large with a rich description of children's early experiences in school.

ECLS-K:1998-99

The database used in this paper to provide priors for the analysis of ECLS-K:2010-11 is the 1998-1999 ECLS-K cohort (NCES, 2001). ECLS-K:1998-99 implemented a multistage probability sample design to obtain a nationally representative sample of children attending kindergarten in 1998-99. The primary sampling units at the base-year of data collection (Fall Kindergarten) were geographic areas consisting of counties or groups of counties. The second stage units were schools within sampled PSUs. The third- and final-stage units were children within schools.

For ECLS-K:1998-99, detailed information about children's kindergarten experiences as well as transition into the formal schooling from Grades 1 through 8 was collected in the fall and the spring of kindergarten (1998-99), the fall and spring of 1st grade (1999-2000), the spring of 3rd grade (2002), the spring of 5th grade (2004), and the spring of 8th grade (2007), seven time points in total. For more detail regarding the sampling design for ECLS-K:1998-99, please see Tourangeau et al. (2009).

[5] Specifically *Pareto-smoothed importance sampling LOO* (PSIS-LOO) is implemented in the loo software to account for the known instability in the loo weights (Vehtari et al., 2017).

ECLS-K:2010-11

As with ECLS-K:1998-99, the ECLS-K:2010-11 is a nationally representative sample of approximately 18,200 children who enrolled in 970 schools during the 2010-11 school year and participated in the base-year of the ECLS-K:2010-11. The children attended both public and private schools. The sample includes children from different racial/ethnic and socioeconomic backgrounds.

As with ECLS-K:1998-99, a multistage sampling design was employed for the ECLS-K:2010-11 cohort. The first stage of national sampling involved the selection of 90 primary sampling units, which consisted of counties and county groups. In the second stage, schools were selected from a sampling frame that was developed based on the NAEP 2010 assessment and came from the NCES 2006 to 2007 Common Core of Data Universe File. Private schools in the NAEP frame were derived from the NCES 2007 to 2008 Private School Survey. In the third stage of sampling, approximately 23 Kindergartners were selected from a list of all enrolled Kindergartners or students of Kindergarten age being educated in an ungraded classroom in each of the sampled schools.

Similar to the ECLS-K:1998-99 study, the children in the ECLS-K:2010-11 comprised a nationally representative sample selected from both public and private schools attending both full-day and part-day Kindergarten Fall of 2010. The children came from diverse socioeconomic and racial/ethnic backgrounds, and the sample includes both children in Kindergarten for the first time and Kindergarten repeaters. Also participating in the study were the children's parents, teachers, schools, and before- and after-school care providers.

The ECLS-K:2010-11 follows the same children from Kindergarten through the fifth grade. Information was collected in the fall and the spring of Kindergarten (2010-11), the fall and spring of first grade (2011-12), the fall and spring of second grade (2012-13), the spring of third grade (2014), the spring of fourth grade (2015), and the spring of fifth grade (2016), 9 time points in total with 6 time points being the same grade levels as those collected in the ECLS-K:1998-99 study (i.e., fall and spring of kindergarten, fall and spring of first grade, the spring of third grade and the spring of fifth grade). Note that although the study refers to later rounds of data collection by the grade the majority of children were expected to be in (that is, the modal grade for children who were in Kindergarten in the 2010-11 school year), children were included in subsequent data collections regardless of their grade level.

Design of the case study

We conducted a comprehensive case study to evaluate the predictive performance of various Bayesian historical borrowing methods via a growth curve model with longitudinal assessments. For the historical cycle ECLS-K:1998-99, we used the data from six time points that are in common with ECLS-K:2010-11, including the fall and spring of kindergarten, the fall and spring of first grade, the spring of third grade and the spring of fifth grade. The spring of eighth grade is not used considering that this time point was not collected for ECLS-K:2010-11.

We evaluated the effects of socio-economic status (SES) at the student level and the percentage of free lunch eligible students (denoted as FLunch) at the school level on the growth trajectory of students' reading scores from fall kindergarten to spring of third

Table 1 Summary statistics for ECLS-K:1998-99 and ECLS-K:2010-11

ECLS-K:1998-99	N	Mean	SD	Range	Skewness	Kurtosis	SE of the Mean
SES	3588	−0.06	0.75	7.11	−0.15	3.37	0.01
FLunch	3588	32.89	25.11	93.00	0.71	−0.33	0.42
Reading at Time 0	3588	34.61	10.20	117.27	3.36	20.60	0.17
Reading at Time 1	3588	45.85	13.70	134.12	2.19	8.24	0.23
Reading at Time 2	3588	51.83	17.27	136.62	2.12	6.39	0.29
Reading at Time 3	3588	75.79	23.66	156.92	0.83	0.69	0.40
Reading at Time 7	3588	125.09	28.33	144.76	−0.16	−0.51	0.47
Reading at Time 11	3588	148.45	26.64	136.95	−0.42	−0.21	0.44
ECLS-K:2010-11	**N**	**Mean**	**SD**	**Range**	**Skewness**	**Kurtosis**	**SE of the Mean**
SES	3904	−0.19	0.81	4.77	0.45	−0.31	0.01
FLunch	3904	52.13	32.49	100.00	−0.06	−1.37	0.52
Reading at Time 0	3904	53.35	11.35	131.19	1.97	7.56	0.18
Reading at Time 1	3904	67.36	13.74	100.30	1.30	3.07	0.22
Reading at Time 2	3904	75.46	16.25	100.58	0.89	0.71	0.26
Reading at Time 3	3904	92.55	17.56	102.66	−0.07	−0.39	0.28
Reading at Time 4	3904	99.25	17.67	115.18	−0.06	−0.24	0.28
Reading at Time 5	3904	110.18	17.36	98.88	−0.38	−0.05	0.28
Reading at Time 7	3904	118.53	15.74	89.56	−0.41	0.04	0.25
Reading at Time 9	3904	127.10	15.24	80.64	−0.65	0.69	0.24
Reading at Time 11	3904	134.16	15.96	85.23	−0.77	0.42	0.26

SES: Socioeconomic status; FLunch: Percentage of students eligible for free lunch; Time 0: Fall of Kindergarten; Time 1: Spring of Kindergarten; Time 2: Fall of 1st Grade; Time 3: Spring of 1st Grade; Time 4: Fall of 2nd Grade; Time 5: Spring of 2nd Grade; Time 7: Spring of 3rd Grade; Time 9: Spring of 4th Grade; Time 11: Spring of 5th Grade

grade and then used the model to predict the reading score of fifth grade for the current cycle of ECLS-K:2010-11. Missing data was addressed by performing multilevel multiple imputation for each cycle separately using the Blimp software program (Enders et al., 2018; Keller & Enders, 2019). For simplicity, we used the first imputed data set for our analyses.[6] Summary statistics for SES and free lunch eligible student percentage as well as for reading scores at different time points for ECLS-K:1998-99 and ECLS-K:2010-11 are shown in Table 1. The time points are coded based on the number of semesters from the fall of kindergarten with the fall kindergarten denoted as time point 0. For example, the spring of kindergarten is one semester away from the fall of kindergarten and thus its time point is 1. For ECLS-K:2010-11, we used the first eight time points (Time 0 to 5, 7 and 9) to estimate growth curve parameters via different historical borrowing methods and then evaluated their performance on predicting students' reading scores at the last time point (Time 11), which is the spring of fifth grade.

Model specification

A Bayesian multilevel growth curve model is fit with reading scores at different time points as level-1, student as level-2, and school as level-3, which is consistent with the design of both cycles. The reading score is the outcome, SES is the student-level

[6] We recognize that it would be optimal to use all the multiply imputed data sets, but evaluating growth trajectories based on multiple imputed data sets is beyond the scope of this paper.

predictor, and percentage of free lunch eligible students is the school-level predictor. The intercept, linear and quadratic terms of time were included in the model to evaluate students' starting points, growth rates and acceleration rates of reading achievement. The interaction between linear and quadratic terms with SES were also included. The student-level intercept and slope (starting point and growth rate) and the school-level intercept were allowed to be random (Bollen & Curran, 2006; Kaplan, 2009).

As the scales of variables included in the models vary greatly, all the variables were standardized first and their z-scores were used in the estimation. Then, all the parameters were converted back to their original scales after the estimation.

Sample size

We evaluated the performance of different priors using the sample of female students only ($N = 1861$). The results for the male students were virtually identical across all conditions of the case study and are available in the supplementary material. To evaluate the impact of sample size on the performance of different Bayesian historical borrowing methods, in addition to the full female sample, a small subsample of female students from high poverty schools was obtained (defined as 75% or above of students in a school who are free lunch eligible). The female subsample has $N = 620$ students.

Choice of priors

We evaluated the performance of dynamic priors, which incorporate the potential heterogeneity between historical data and current data through a joint prior distribution, and compared it to regular priors with predetermined prior values and strength. Specifically, for dynamic priors, we varied the IG prior for τ^2 at IG (1, 1) and IG (1, 0.001) to allow for different degrees of borrowing for coefficients. Moreover, the precision matrix of the random intercept and random slope has a Wishart distribution prior[7], $\mathrm{W}(\nu, \nu\mathbf{S}^{-1})$ where ν takes on the values 2 (weak borrowing) or 20 (strong borrowing) and $\mathbf{S} = \mathbf{\Sigma}_S'\mathbf{\Omega}\mathbf{\Sigma}_S$ is the baseline covariance matrix where $\mathbf{\Sigma}_S$ is a diagonal matrix whose diagonal elements are distributed as half-Cauchy(0,1) and $\mathbf{\Omega} \sim$ LKJCorr(1) (Lewandowski et al., 2009). For commensurate priors, we also used a Wishart prior for the precision matrix of the random intercept and random slope, $\mathrm{W}(\nu, \nu\mathbf{S}^{-1})$, where ν takes on the value of 2. For power priors, we varied the a parameter using values of .25, .50 and .75. For aggregated data-dependent priors, the estimated coefficients from historical data were used as the prior mean and the prior variances of historical coefficients were used as the prior variances. For comparison purposes, two extreme kinds of borrowing, complete pooling and no borrowing of the historical data sets were also examined. We specified a weakly informative half-Cauchy (0,1) prior for the standard deviation σ of the individual-level error term, and a non-informative $N(0, 10^2)$ prior for the school-level coefficients across all cycles ($\mathbf{\Gamma}^0$) in BDB and the mean school-level coefficients in the current cycle ($\boldsymbol{\mu}$) in the non-informative prior conditions. All analyses were conducted within the R

[7] Note that we utilized the Wishart prior for the student-level precision matrix in both the case study and the simulation study as it demonstrated better convergence properties compared to using the inverse-Wishart distribution. We then scaled the results back to a covariance matrix.

programming environment (R Core Team, 2022) using rstan (Stan Development Team, 2021). All code for the case study are available in the supplementary material.

Results of the case study

As mentioned earlier, there are two different scenarios evaluated in the case study, namely, female students in high poverty schools and the full sample of female students. For these scenarios, results are presented in two tables with the first for regression coefficient and variation estimates and the second for predictive performance. These results are presented in Tables 2 and 3 for female students in high poverty schools, and Tables 4 and 5 for all female students, respectively.

Across different borrowing methods, Bayesian dynamic borrowing and commensurate priors provided similar coefficient and variation estimates to those with no borrowing, indicating that the historical data and current data are heterogeneous. Pooling and power priors with greater amounts of borrowing (i.e., $a = 0.5$ and $a = 0.75$), on the other hand, provided similar coefficient and variation estimates and showed differences from no borrowing and dynamic borrowing priors, particularly on the coefficient and variation estimates of student-level and school-level intercepts.

In terms of predictive performance, overall, different borrowing methods performed similarly for full samples and high poverty school samples. Theil's equality coefficients were nearly identical, with pooling and power priors having slightly smaller variance proportions. Pooling, aggregated data-dependent priors, and power priors had relatively smaller RMSE between predicted and observed reading scores at the spring of 5th grade and smaller KLD. No borrowing provided larger RMSE and KLD, but the smallest LOOICs. Bayesian dynamic borrowing under IG(1, 0.001) had a smaller LOOIC compared to pooling and power priors. Across all the prediction evaluation criteria, aggregated data-dependent priors performed well overall (i.e., smaller RMSE and KLD compared to no borrowing and smaller LOOICs compared to pooling, power priors and dynamic borrowing methods). Due to the heterogeneity between the historical data and the current data as reflected in Table 1, Bayesian dynamic borrowing methods did not outperform other borrowing methods, but would still be a reasonable choice in terms of providing smaller LOOICs compared to pooling and smaller RMSEs compared to no borrowing.

Design of the simulation study

The results of the case study indicate that the cycles of ECLS-K:1998-99 and ECLS-K:2010-11 are relatively heterogeneous in terms of the effects we evaluated such that Bayesian dynamic borrowing and commensurate priors borrow less due to data heterogeneity and provide estimates similar to Bayesian multilevel growth curve model with non-informative priors (i.e., no borrowing). In order to study the performance of different Bayesian historical borrowing methods under different levels of data heterogeneity as well as varying levels of sample size, a comprehensive simulation study was further conducted.

Model specification and estimation

For the simulation study, a Bayesian multilevel linear growth curve model was used (details to follow). For the Markov chain Monte Carlo simulations, 2000 iterations

Table 2 Growth curve model coefficient and variation estimates for the sample of female students in high poverty schools

	Regression coefficients							Stu-Level			Sch-level	
	Intcp	*t*	t^2	SES	FLunch	*t* × SES	t^2 × SES	Var(Intcp)	Cov(Intcp,*t*)	Var(*t*)	Var(Intcp)	Var(R)
BLR Noninformative	47.75	15.00	− 0.74	3.94	0.05	1.20	− 0.11	87.64	1.30	1.29	7.68	47.91
BLR Pooling	32.90	14.52	− 0.61	3.37	0.17	1.04	− 0.05	87.61	0.53	2.85	105.69	55.52
BLR AGDP	50.64	13.73	− 0.61	3.82	0.03	0.30	− 0.01	88.29	1.32	1.28	7.45	48.58
PP (0.25)	34.85	14.83	− 0.67	3.60	0.15	1.13	− 0.07	84.13	1.67	1.85	76.06	54.23
PP (0.5)	33.31	14.76	− 0.64	3.44	0.16	1.11	− 0.06	85.36	1.12	2.46	95.75	54.49
PP (0.75)	33.08	14.64	− 0.62	3.39	0.16	1.07	− 0.06	86.75	0.76	2.72	102.90	55.00
BDB IG(1,1) W2,W2	47.64	15.00	− 0.74	3.94	0.05	1.20	− 0.11	85.52	1.69	1.19	7.23	54.22
BDB IG(1,0.001) W2,W2	48.11	14.94	− 0.73	3.88	0.05	1.14	− 0.10	85.35	1.69	1.20	7.47	54.22
BDB IG(1,1) W2,W20	47.48	15.01	− 0.74	3.93	0.05	1.20	− 0.11	85.20	1.71	1.19	7.32	54.25
BDB IG(1,.001) W2,W20	47.75	14.95	− 0.74	3.82	0.05	1.15	− 0.10	85.28	1.70	1.19	7.24	54.23
BDB IG(1,1) W20,W2	47.60	15.00	− 0.74	3.94	0.05	1.20	− 0.11	84.96	1.70	1.23	7.26	54.20
BDB IG(1,.001) W20,W2	47.76	14.94	− 0.73	3.88	0.05	1.14	− 0.10	84.91	1.70	1.23	7.64	54.19
BDB IG(1,1) W20,W20	47.60	14.99	− 0.74	3.92	0.05	1.19	− 0.11	85.22	1.68	1.23	6.95	54.20
BDB IG(1,.001) W20,W20	48.02	14.94	− 0.73	3.87	0.05	1.14	− 0.10	85.33	1.67	1.23	6.87	54.16
CP W2,W2	48.12	14.96	− 0.74	3.81	0.05	1.17	− 0.10	85.02	1.74	1.18	6.58	54.25

Intcp:intercept; *t*: time; Stu-Level: student-level; Sch-Level: school-level; Var(R): variance of time-level residual; BLR: Bayesian Linear Regression; AGDP: aggregated data-dependent prior; PP: power prior; BDB: Bayesian dynamic borrowing; IG: inverse-gamma prior for time-level variance of the joint prior distribution, which determines the degree of time-level borrowing; W2: Wishart prior with weak borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); W20: Wishart prior with strong borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); CP: commensurate prior

Table 3 Prediction performance of different borrowing methods for the sample of female students in high poverty schools

	Theil inequ coef	Theil bias prop	Theil var prop	RMSE	KLD	LOOIC (Student)	LOOIC (School)
BLR Noninformative	0.03	0.01	0.11	9.93	0.07	33967.63	33621.26
BLR Pooling	0.03	0.01	0.06	8.72	0.02	34192.96	33837.50
BLR AGDP	0.03	0.01	0.12	8.81	0.03	34027.67	33692.09
PP (0.25)	0.03	0.01	0.09	8.91	0.03	34124.25	33794.19
PP (0.5)	0.03	0.01	0.07	8.83	0.02	34145.22	33802.78
PP (0.75)	0.03	0.01	0.06	8.78	0.02	34191.10	33838.18
BDB IG(1,1) W2,W2	0.03	0.01	0.13	9.89	0.07	34090.48	33779.65
BDB IG(1,0.001) W2,W2	0.03	0.01	0.13	9.83	0.07	34093.57	33788.90
BDB IG(1,1) W2,W20	0.03	0.01	0.13	9.90	0.07	34086.56	33796.26
BDB IG(1,0.001) W2,W20	0.03	0.01	0.13	9.85	0.07	34089.09	33784.94
BDB IG(1,1) W20,W2	0.03	0.01	0.13	9.90	0.07	34107.72	33780.68
BDB IG(1,0.001) W20,W2	0.03	0.01	0.13	9.84	0.07	34097.21	33777.34
BDB IG(1,1) W20,W20	0.03	0.01	0.13	9.91	0.07	34085.33	33784.04
BDB IG(1,0.001) W20,W20	0.03	0.01	0.13	9.84	0.07	34084.31	33770.53
CP W2,W2	0.03	0.01	0.13	9.84	0.07	34087.65	33786.41

Theil Inequ Coef: Theil's inequaility coefficient; Theil Bias Prop: Theil's bias proportion; Theil Var Prop: Theil's variance proportion; RMSE: root mean squared error of the predicted reading score vs. observed reading score at the spring of 5th grade; KLD: Kullback-Leibler Divergence Score; LOOIC (Student): Leave-one-out cross validation information criterion with each student left out one at a time; LOOIC (School): Leave-one-out cross validation information criterion with each school left out one at a time; BLR: Bayesian Linear Regression; AGDP: aggregated data-dependent prior; PP: power prior; BDB: Bayesian dynamic borrowing; IG: inverse-gamma prior for time-level variance of the joint prior distribution, which determines the degree of time-level borrowing; W2: Wishart prior with weak borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); W20: Wishart prior with strong borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); CP: commensurate prior

(where the first 1000 were used as warm-up iterations and discarded) were run for each of the four chains. We ran 500 replications, out of which the model converged well for at least 489 replications. Only the replications with converged models were used. The replications with split-chain potential scale reduction factors $\hat{R} \geq 1.05$ were discarded (see e.g. Gelman et al., 2014).

Data generation and simulation conditions

For our simulation study, we evaluated the impact of the number of schools, school size, heterogeneity of historical information, and prior choice. We used the historical cycle ECLS-K:1998-99 as the base to generate the data of the current cycle. Let G denote the number of schools and let n denote the number of students per school. We examined four different sample sizes: (1) $G = 10$, $n = 20$; (2) $G = 10$, $n = 40$, (3) $G = 20$, $n = 20$,

Table 4 Growth curve model coefficient and variation estimates for the full sample of female students

	Regression coefficients							Stu-level			Sch-level	
	Intcp	*t*	t^2	SES	FLunch	*t* × SES	t^2 × SES	Var(Intcp)	Cov(Intcp,*t*)	Var(*t*)	Var(Intcp)	Var(R)
BLR Noninformative	55.83	15.49	− 0.79	4.65	− 0.04	1.24	− 0.12	120.90	− 2.29	1.28	8.27	53.47
BLR Pooling	41.20	15.91	− 0.61	3.88	0.05	1.93	− 0.12	140.88	− 9.61	6.05	102.35	83.57
BLR AGDP	59.22	13.02	− 0.53	3.06	− 0.04	1.13	− 0.10	120.75	− 2.17	1.23	8.88	58.73
PP (0.25)	43.09	15.92	− 0.65	4.36	0.03	1.59	− 0.10	120.31	− 7.16	4.51	84.96	82.60
PP (0.5)	41.70	16.01	− 0.63	4.11	0.04	1.77	− 0.11	130.23	− 8.45	5.56	98.41	83.05
PP (0.75)	41.28	15.96	− 0.62	4.00	0.05	1.86	− 0.11	136.78	− 9.16	5.89	101.23	83.37
BDB IG(1,1) W2,W2	55.83	15.49	− 0.79	4.66	− 0.04	1.23	− 0.12	110.59	− 0.59	0.85	8.56	81.68
BDB IG(1,0.001) W2,W2	55.84	15.51	− 0.79	4.52	− 0.04	1.29	− 0.13	110.65	− 0.58	0.84	8.32	81.68
BDB IG(1,1) W2,W20	55.83	15.49	− 0.79	4.67	− 0.04	1.24	− 0.12	110.13	− 0.58	0.85	9.35	81.68
BDB IG(1,0.001) W2,W20	55.84	15.51	− 0.79	4.52	− 0.04	1.29	− 0.13	110.11	− 0.59	0.85	9.49	81.65
BDB IG(1,1) W20,W2	55.81	15.49	− 0.79	4.67	− 0.04	1.24	− 0.12	110.65	− 0.63	0.88	8.60	81.63
BDB IG(1,0.001) W20,W2	55.83	15.51	− 0.79	4.53	− 0.04	1.29	− 0.13	110.98	− 0.68	0.88	8.50	81.62
BDB IG(1,1) W20,W20	55.83	15.49	− 0.79	4.66	− 0.04	1.23	− 0.12	110.34	− 0.65	0.88	9.43	81.62
BDB IG(1,0.001) W20,W20	55.82	15.51	− 0.79	4.52	− 0.04	1.29	− 0.13	110.22	− 0.64	0.87	9.44	81.64
CP W2,W2	55.98	15.51	− 0.79	4.45	− 0.04	1.29	− 0.13	110.85	− 0.58	0.84	7.63	81.72

Intcp:intercept; *t*: time; Stu-Level: student-level; Sch-Level: school-level; Var(R): variance of time-level residual; BLR: Bayesian Linear Regression; AGDP: aggregated data-dependent prior; PP: power prior; BDB: Bayesian dynamic borrowing; IG: inverse-gamma prior for time-level variance of the joint prior distribution, which determines the degree of time-level borrowing; W2: Wishart prior with weak borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); W20: Wishart prior with strong borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); CP: commensurate prior

Table 5 Prediction performance of different borrowing methods for the full sample of female students

	Theil inequ coef	Theil bias prop	Theil var prop	RMSE	KLD	LOOIC (Student)	LOOIC (School)
BLR Noninformative	0.03	0.01	0.11	10.99	0.17	103859.87	102848.58
BLR Pooling	0.03	0.02	0.04	8.84	0.06	107146.03	105905.94
BLR AGDP	0.03	0.01	0.13	7.86	0.04	105155.18	104154.24
PP (0.25)	0.03	0.02	0.05	8.48	0.05	106807.01	105608.49
PP (0.5)	0.03	0.02	0.05	8.67	0.05	107018.84	105753.76
PP (0.75)	0.03	0.02	0.04	8.76	0.06	107087.27	105811.26
BDB IG(1,1) W2,W2	0.03	0.02	0.18	10.89	0.21	105698.48	104927.28
BDB IG(1,0.001) W2,W2	0.03	0.02	0.18	10.93	0.21	105682.55	104901.07
BDB IG(1,1) W2,W20	0.03	0.02	0.18	10.89	0.21	105675.18	104894.27
BDB IG(1,0.001) W2,W20	0.03	0.02	0.18	10.92	0.21	105670.24	104912.10
BDB IG(1,1) W20,W2	0.03	0.02	0.17	10.89	0.20	105690.84	104929.72
BDB IG(1,0.001) W20,W2	0.03	0.02	0.17	10.91	0.21	105674.95	104929.79
BDB IG(1,1) W20,W20	0.03	0.02	0.17	10.88	0.20	105686.13	104938.19
BDB IG(1,0.001) W20,W20	0.03	0.02	0.17	10.92	0.21	105663.53	104920.70
CP W2,W2	0.03	0.02	0.18	10.92	0.21	105761.37	104970.18

Theil Inequ Coef: Theil's inequaility coefficient; Theil Bias Prop: Theil's bias proportion; Theil Var Prop: Theil's variance proportion; RMSE: root mean squared error of the predicted reading score vs. observed reading score at the spring of 5^{th} grade; KLD: Kullback-Leibler Divergence Score; LOOIC (Student): Leave-one-out cross validation information criterion with each student left out one at a time; LOOIC (School): Leave-one-out cross validation information criterion with each school left out one at a time; BLR: Bayesian Linear Regression; AGDP: aggregated data-dependent prior; PP: power prior; BDB: Bayesian dynamic borrowing; IG: inverse-gamma prior for time-level variance of the joint prior distribution, which determines the degree of time-level borrowing; W2: Wishart prior with weak borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); W20: Wishart prior with strong borrowing for student-level (the former) or school-level (the latter) precision matrix (results were converted back the covariance matrix); CP: commensurate prior

and (4) $G = 20$, $n = 40$. For the historical data ECLS-K:1998-99, a random sample stratified by schools was selected with one of the sample size scenarios mentioned above. The selected within-student variables included the reading score and a linear (t) and quadratic (t^2) component of time. Same as the case study, the between-student variable in the simulation study is SES, and the school-level variable is percentage of students who are eligible for free lunch in each school.

Data for the current cycle with each of the above four sample sizes were generated with different degrees of heterogeneity compared to the historical data. A Bayesian multilevel linear growth curve model was fit on the ECLS-K:1998-99 data with the reading score as the outcome, t and t^2 as within-student predictors, SES as the student-level predictor, and the percentage of free lunch eligible students in each school as the school-level predictor. The interaction terms between the linear and quadratic components of

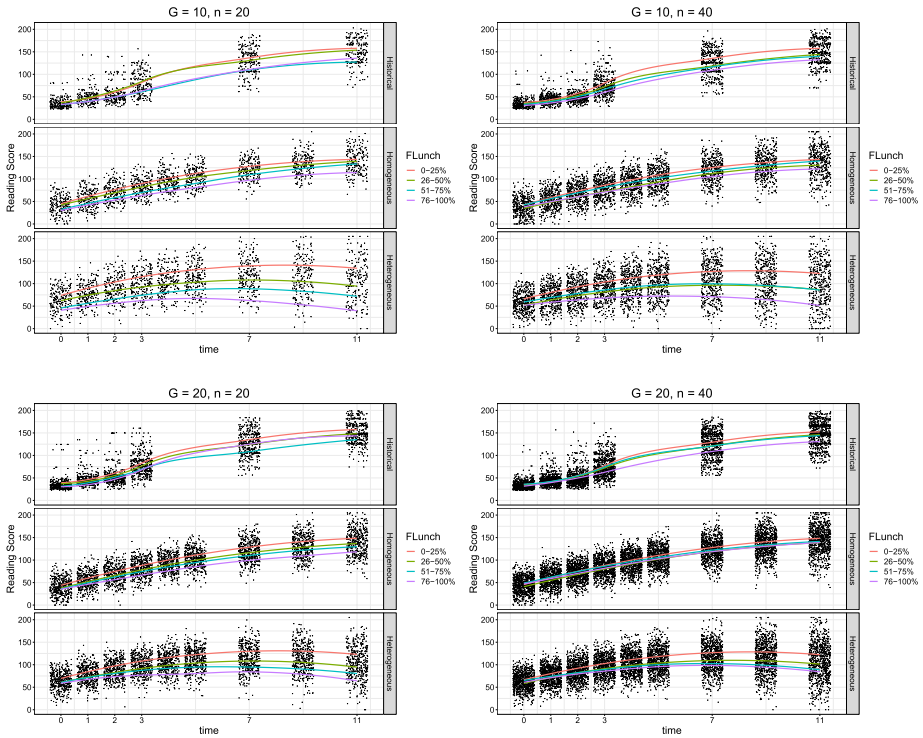

Fig. 1 Growth trajectories of one simulated data set for different sample size and heterogeneity conditions

time and student-level SES as well as school-level free lunch eligible percentage were also included. The student-level intercept and slope, which are students' starting points and growth rates, were allowed to be random. The school-level intercept was treated as random. Fixed effect and random effect estimates from the historical cycle were obtained and used to generate the current cycle's data. That is, for the current cycle, predictor values were sampled from the ECLS-K:2010-11 data with a certain sample size, while the outcome (reading score) was generated with different generating coefficients. For the homogeneous condition, the regression coefficient estimates obtained based on the historical cycle were used as the generating coefficients to generate data for the current cycle. For the heterogeneous condition, the historical regression coefficient estimates with adjustments ranging from − 10% to + 150% were used to generate the data of the current cycles so that the regression coefficients in the current cycle would be heterogeneous compared to the historical regression coefficients (specific adjustments are included in the supplemental material). The growth trajectories of one simulated data set are illustrated in Fig. 1 for historical cycle (upper), homogeneous condition of the current cycle (middle) and heterogeneous condition of the current cycle (bottom), one plot per sample size. Data for the current cycle were generated with the same proportion of schools in each of the four poverty categories differentiated by percentage of students who are eligible for free lunch.

Regarding prediction, we used the same method as discussed in the case study. That is, for the current cycle, we used the first eight time points (Time 0 to 5, 7 and 9) to estimate growth curve parameters via different historical borrowing methods and then

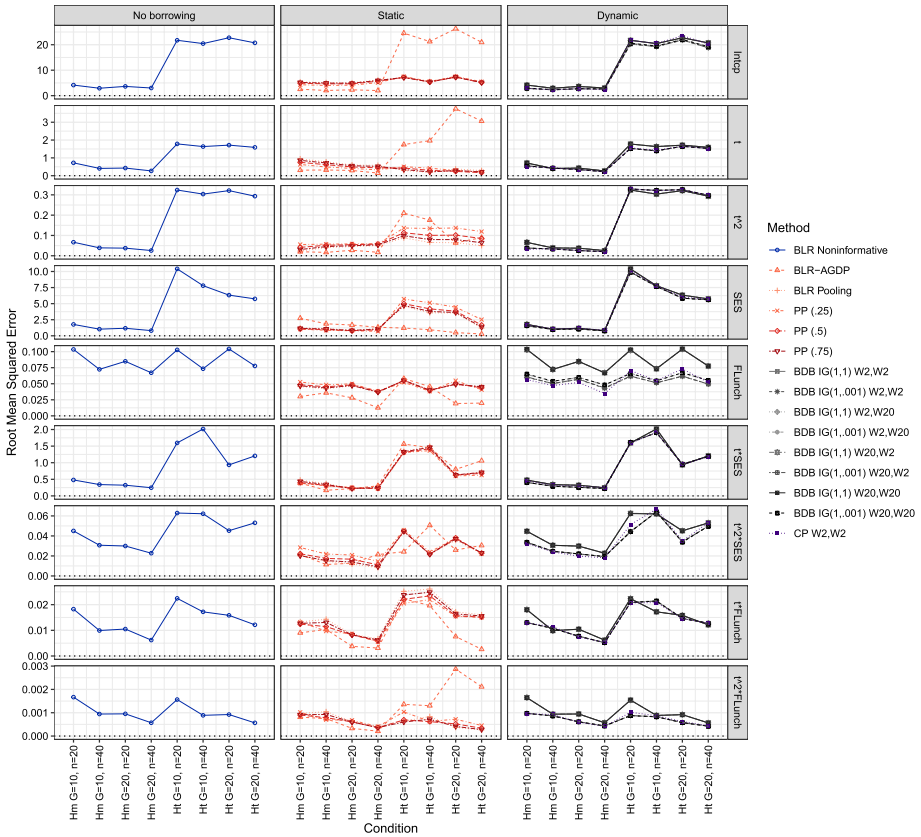

Fig. 2 Root Mean Squared Error (RMSE) of regression coefficient estimates for different sample size and heterogeneity conditions

evaluated their performance on predicting students' reading scores at the last time point (Time 11), which is the spring of fifth grade.

With regard to prior choice, similar to the case study, we assessed the performance of dynamic priors and compared it to the no borrowing case, aggregated data-dependent priors, complete pooling, and power priors. Specifically, for Bayesian dynamic borrowing, we varied the inverse-Gamma prior for τ^2 at IG(1, 1) and IG(1, 0.001). The precision matrices of the student-level random intercepts and random slopes have Wishart distribution $\mathrm{W}(\nu, \nu\mathbf{S}^{-1})$ where ν takes 2 (weak borrowing) or 20 (strong borrowing) and $\mathbf{S} = \mathbf{\Sigma}_S'\mathbf{\Omega}\mathbf{\Sigma}_S$ is the baseline precision where $\mathbf{\Sigma_S}$ is a diagonal matrix whose diagonal elements are distributed as half-Cauchy(0, 1) and $\mathbf{\Omega} \sim$ LKJCorr(1). For commensurate priors, Wishart prior with $\nu = 2$ was used for both student-level and school-level random effects. For power priors, again, we varied a^h at 0.25, 0.50 and 0.75. All code for the simulation study are available in the supplementary material.

Results of the simulation study

Two sets of results from the simulation study are presented, one being the evaluation of growth curve model parameter recovery as illustrated in Fig. 2 (for regression coefficient estimates) and Fig. 3 (for variation estimates), and the other being the quality of prediction on students' reading score at the spring of 5$^{\text{th}}$ grade across different borrowing

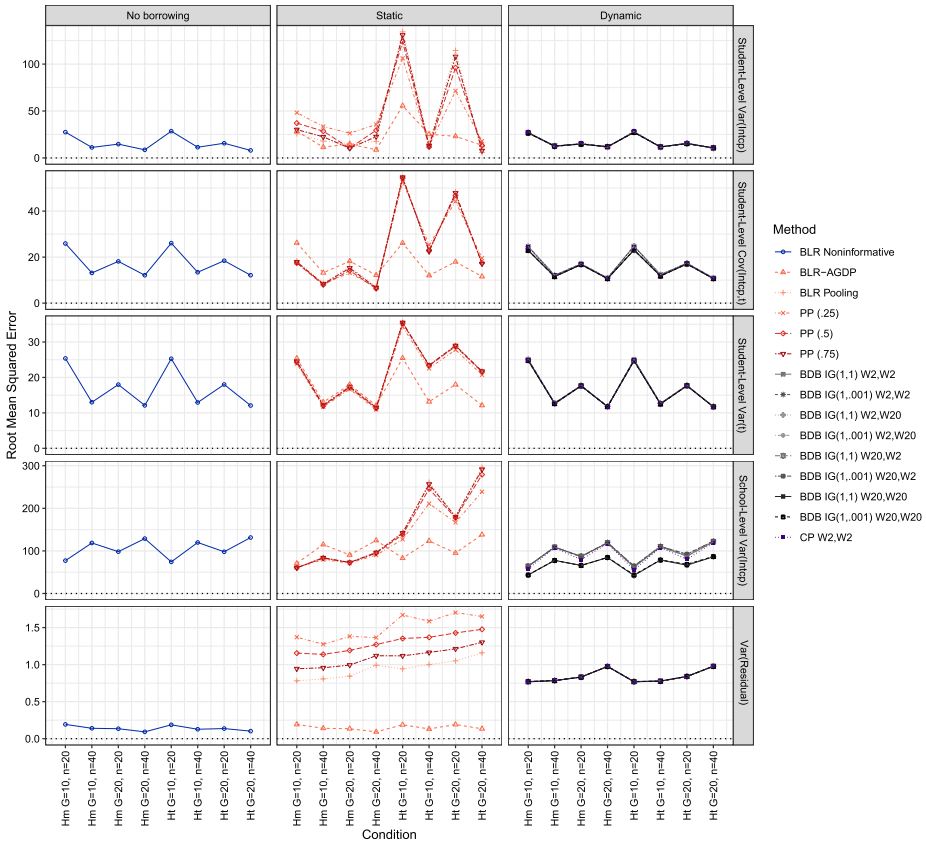

Fig. 3 Root Mean Squared Error (RMSE) of variation estimates for different sample size and heterogeneity conditions

methods as illustrated in Fig. 4 (for the homogeneous condition) and Fig. 5 (for the heterogeneous condition).

Figures 2 and 3 display three columns of plots indicating no borrowing, static borrowing (pooling, aggregated data-dependent prior, and power prior), and dynamic borrowing (BDB and commensurate prior). The x-axis indicates two different heterogeneity (hm—homogeneous; ht—heterogeneous) by four different sample size conditions (G—number of schools; n—number of students per school), in total eight different conditions. The rows indicate different regression coefficients (in Fig. 2) or random variations (in Fig. 3), including student-level variance and covariance of random intercept and slopes, school-level variance of random intercept, and variance of residuals.

Root mean squared errors (RMSEs) between the generating/true parameters and estimated parameters are used to evaluate the accuracy of parameter estimates by different borrowing methods. Figure 2 shows that overall, RMSEs for regression coefficient estimates were smaller when sample sizes were larger and when the current data was more homogeneous to the historical data. Under the same sample size and heterogeneity condition, dynamic borrowing methods were better or similar to no borrowing across different regression coefficients. For example, BDB under the IG (1, 0.001) hyperprior and commensurate prior provided smaller RMSEs for the coefficient of FLunch than no borrowing. The static borrowing methods provided similar RMSEs to those with no

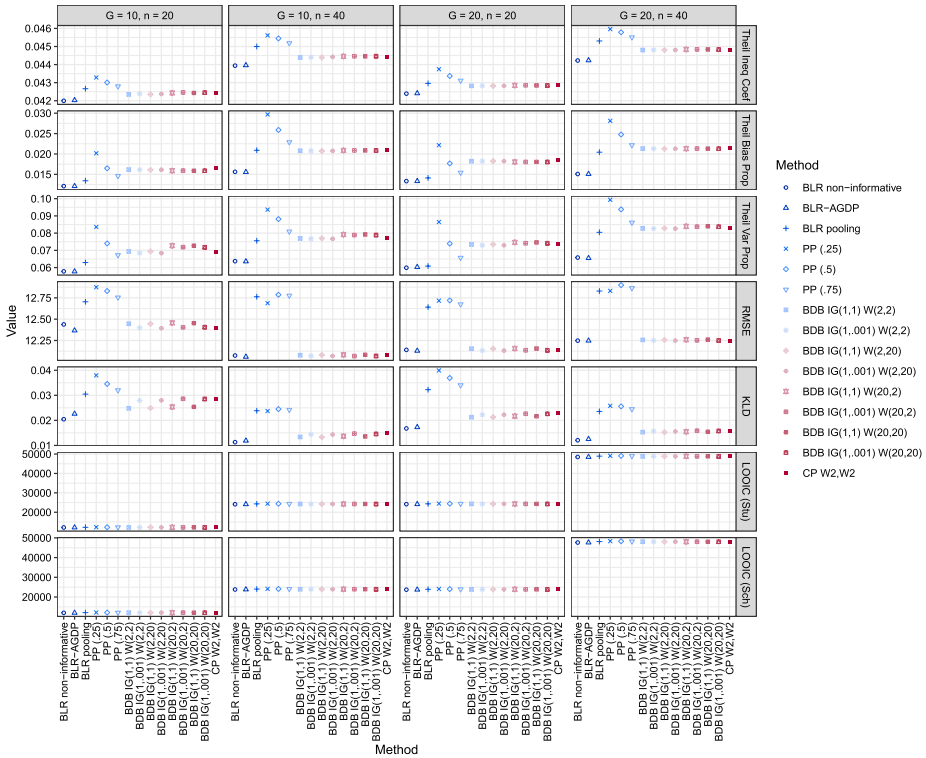

Fig. 4 Prediction performance of different borrowing methods under the homogeneous condition

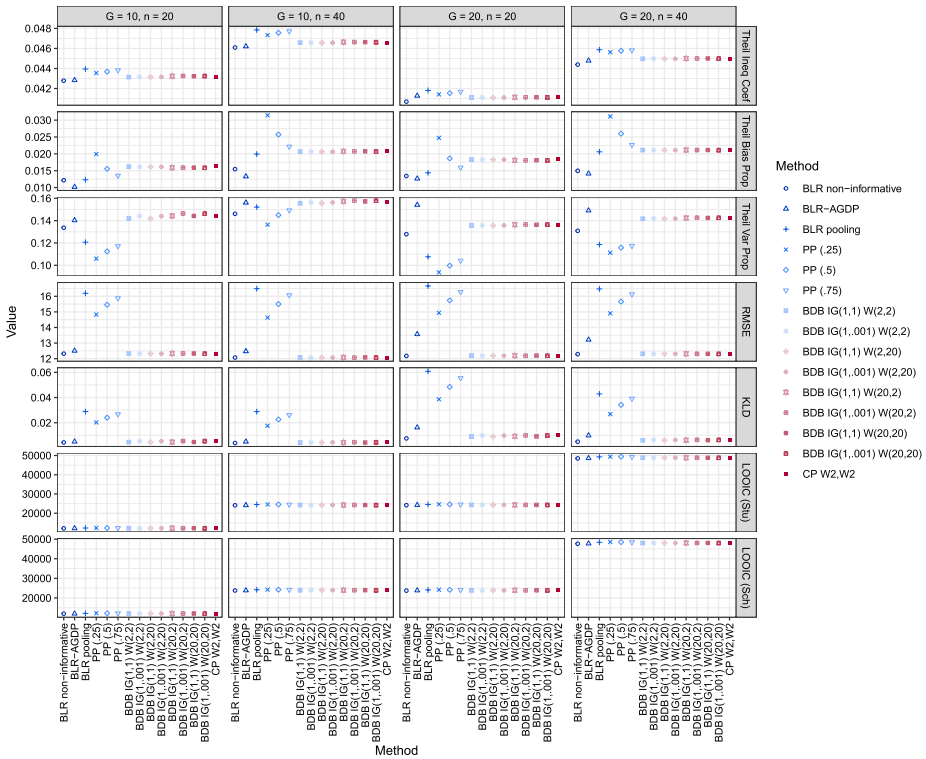

Fig. 5 Prediction performance of different borrowing methods under the heterogeneous condition

borrowing and dynamic borrowing methods under the homogeneous condition, but the performance varied under the heterogeneous condition. Specifically, under the heterogeneous condition, the aggregated data-dependent priors provided the smallest RMSE for the quadratic term of time (acceleration rate), but the largest RMSEs for the intercept (starting point) and the linear term of time (growth rate). The power prior methods provided the same or better RMSEs for the regression coefficient estimates overall.

For the variation estimates shown in Fig. 3, no borrowing, aggregated data-dependent priors, and dynamic borrowing provided similar RMSEs for student-level random variation estimates. Nevertheless, BDB under IG(1, 0.001) and Wishart(20) hyperpriors provided the smallest RMSE for school-level random intercept variation. No borrowing and AGDP provided the smallest RMSEs for the variance of residuals. Power priors, in contrast, provided the largest RMSEs for the variance of residuals. Under the heterogeneous condition, power priors yielded the largest RMSEs for all the variation estimates.

In terms of predictive performance, the Theil inequality coefficients, bias and variance proportions, RMSE between the observed and predicted reading score at the spring of 5th grade, KLD, LOOIC with one student left out at a time and LOOIC with one school left out at a time were adopted to evaluate the prediction quality of different borrowing methods under different sample size and heterogeneity conditions. As Fig. 4 shows, when the current data and the historical data were relatively homogeneous, we found that no borrowing, AGDP, and dynamic borrowing performed similarly in terms of prediction. There were minor differences in Theil's inequality coefficients, bias proportion and variance proportion, where no borrowing and AGDP performed slightly better. Based on RMSE for prediction and KLD, power priors performed worse than other borrowing methods. Pooling was slightly better than power priors, but also did not outperform no borrowing, AGDP, or dynamic borrowing methods. Across different sample sizes, the predictive performance based on the LOOIC across different borrowing methods was similar.

When the current data and the historical data were heterogeneous, as Fig. 5 shows, pooling performed worse on prediction compared to its performance under the homogeneous condition and had the largest RMSE of prediction and KLD. Power priors also had large RMSE of prediction and KLD compared to no borrowing, AGDP and dynamic borrowing. BDB and commensurate priors performed similarly to no borrowing and AGDP in terms of prediction. Similar to the homogeneous condition, the predictive performance based on the LOOIC across different sample size conditions was similar.

Conclusion

As we noted in the introduction, longitudinal data are becoming increasingly available and are powerful data collection strategies that allow for the study of important outcomes in education and the social and behavioral sciences. That said, longitudinal data are also extremely expensive to collect, and except in rare cases, few longitudinal studies exist that have gone beyond two separate cohorts. Thus, it becomes even more critical that new methodologies be developed to leverage information across longitudinal studies while accounting for the heterogeneity that can be induced by cohort effects as well as other changes in the data collection strategies across cycles.

The purpose of this study was to build on recent work by Kaplan et al. (2022), and examine a variety of methods under the umbrella of Bayesian historical borrowing with a specific focus on models for growth in longitudinal studies. Outcomes of interest in our study were the parameter recovery and predictive performance of growth models under a variety of historical borrowing methods. We utilized both a case study and a simulation study in the context of one historical cycle insofar as this is a relatively realistic situation. We note that this is in contrast to Kaplan et al. (2022), who examined Bayesian historical borrowing for cross-sectional multilevel models based on the structure of PISA and with five historical cycles.

Our findings show that in the case of one historical cycle, most methods of historical borrowing perform similarly with respect to predictive performance and parameter recovery. We note that in the present paper, and consistent with Kaplan et al. (2022), pooling and power priors performed relatively poorly across the conditions in this study, particularly when the current data and the historical data were heterogeneous. The findings regarding power priors are not surprising insofar as previous studies have shown relatively poor performance of power priors in a variety of settings (see Du et al., 2020, and references therein), though power priors were not examined in the case of longitudinal studies with a focus on prediction. The findings regarding pooling under the homogeneous condition are a bit surprising insofar as the homogeneous condition of our simulation study mimicked the desirable conditions for combining data from longitudinal studies as described in Hofer and Piccinin (2009), e.g. common time points and identical measurements. However, the current data and the historical data might still not have been homogeneous enough given that the generating model is a complex multilevel growth curve model. Overall though, the findings show that using aggregated data-dependent priors or simply using the current cycle of data with non-informative priors performed well. We speculate that this is because we only examined the realistic condition of one historical study. Note, however, that Kaplan et al. (2022) found clear benefits of Bayesian dynamic borrowing over other methods of historical borrowing in the cross-sectional multilevel case study using PISA (OECD, 2002; 2019) with five historical cycles, and so, in line with Kaplan et al. (2022), we argue that Bayesian dynamic borrowing or commensurate priors is a prudent choice for borrowing information from previous cycles of data and that sensitivity analyses examining a variety of borrowing procedures to gauge the extent of homogeneity or heterogeneity across data sets would be advisable. Also, future research should expand the current study to consider more cycles of longitudinal data.[8]

It is important to point out some limitations of this study as they pertain specifically to the application of these methods to longitudinal data. First, we did not account for the sampling weights, which are critical in large-scale assessments. The problem of sample weighting in Bayesian models generally has been discussed in Gelman (2007), who declared at the time that "Survey weighting is a mess" (pg. 153). Since then, however, there have been important developments in the implementation of sampling weights for Bayesian analyses with applications to problems of survey data (see e.g. Goldstein, 2011;

[8] As of this writing, the United States National Center for Education Statistics in the process of designing and launching the Early Childhood Longitudinal Study: Kindergarten Class of 2023-24.

Trendtel & Robitzsch, 2021). Another approach advocated by Gelman (2007) would be multilevel regression with post-stratification (Gelman & Thomas, 1997). This approach is mostly impractical in the context of large-scale assessments insofar population counts would be required for all variables used in an analysis, and such counts are typically only available for demographic variables. In any case, both approaches would require considerable future research, which was beyond the scope of this paper. Second, our paper suffers from limitations common to methodological research—namely that real data can not be expected to mimic ideal conditions found in simulation studies, and simulation studies can not examine all possible conditions that would be encountered in real data scenarios. Nevertheless, on the basis of the findings in this paper, we conclude that Bayesian historical borrowing methods should be given serious consideration, and that a comparison of methods based on in-sample and pseudo out-of-sample performance measures should be a routine part of the workflow when multiple cycles of longitudinal data are available from which to borrow from and when prediction is of central focus.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-022-00140-w.

Additional file 1. Bayesian Historical Borrowing with Longitudinal Large-Scale Assessments.

Acknowledgements
The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D190053 to The University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.
This research was performed using the computing resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.
The authors are grateful to an anonymous reviewer who made valuable comments on a previous draft of this paper.

Author contributions
DK conceptualized the paper and serves as the Principal Investigator of the project. DK also contributed to the writing of the paper. JC contributed to the design, analysis, and writing of the paper and serves as the co-Principal Investigator of the project. WL contributed to the analysis of the case study and simulation study and writing of the paper. SY contributed to the analysis of the case study and simulation study and tables/plots of the paper. All authors read and approved the final manuscript.

Funding
The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D190053 to The University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. This research was performed using the computing resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

Availability of data and materials
This project made use of public use files of the Early Childhood Longitudinal Study for the case study and simulated data for the simulation studies. All data are freely available, and starting seeds are available for the simulation studies.

Declarations

Ethics approval and consent to participate
This research was based on secondary analyses of public-use databases and simulated dat. No ethics approval was required and no human subjects were involved in the research.

Consent for publication
The authors, DK, JC, WL, and SY provide consent for publication of this paper in the journal.

Competing interests
There are no competing interests.

Received: 30 March 2022 Accepted: 9 November 2022
Published online: 18 January 2023

References

Bainter, S. A., & Curran, P. J. (2015). Advantages of integrative data analysis for developmental research. *Journal of Cognition and Development, 16*(1), 1–10.
Bernardo, J., & Smith, A. F. M. (2000). *Bayesian theory*. Wiley.
Betancourt, M. (2018). Bayes sparse regression. (https://betanalpha.github.io/assets/case studies/bayes sparse regression.html, Last accessed: 2022-02-27)
Blossfeld, H.-P., & Roßbach, H.-G. E. (2019). Education as a lifelong process: The German national educational panel study (neps) (2nd ed.). Springer
Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. John Wiley & Sons.
Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97,* 465–480. https://doi.org/10.1093/biomet/asq017.
Chen, M.-H., Ibrahim, J. G., & Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference, 84,* 121–137.
Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14,* 81–100.
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.
Du, H., Bradbury, T. N., Lavner, J. A., Meltzer, A. L., McNulty, J. K., Neff, L. A., & Karney, B. R. (2020). A comparison of Bayesian synthesis approaches for studies comparing two means: A tutorial. *Research Synthesis Methods, 11,* 36–65. https://doi.org/10.1002/jrsm.1365.
Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23*(2), 298–317.
Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515–533.
Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153–164.
Gelman, A., Carlin, J. B., Stern, D. B., Dunson, H. S., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). Chapman & Hall.
Gelman, A., & Thomas, L. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23,* 127–135.
Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102,* 359–378.
Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Wiley.
Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., & Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics, 67,* 1047–1056.
Hobbs, B. P., Carlin, B. P., & Sargent, D. J. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis, 7*(2), 1–36.
Hofer, S., & Piccinin, A. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14,* 150–64. https://doi.org/10.1037/a0015566.
Ibrahim, J. G., Chen, M.-C., Gwon, Y., & Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine*. https://doi.org/10.1002/sim.6728.
Ibrahim, J. G., & Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science, 15,* 46–60.
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.). Sage Publications.
Kaplan, D., Chen, J., Yavuz, S., & Lyu, W. (2022). Bayesian dynamic borrowing of historical information with applications to the analysis of large-scale assessments. *Psychometrika*. https://doi.org/10.1007/s11336-022-09869-3.
Kaplan, D., & George, R. (1998). Evaluating latent variable growth models through ex post simulation. *Journal of Educational and Behavioral Statistics, 23,* 216–235.
Kaplan, D., & Huang, M. (2021). Bayesian probabilistic forecasting with state NAEP data. *Large-Scale Assessements in Education*. https://doi.org/10.1186/s40536-021-00108-2.
Keller, B. T., & Enders, C. K. (2019). Blimp user's guide (version 2.1).
Kullback, S. (1959). *Information theory and statistics*. New York: John Wiley and Sons.
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.
Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100,* 1989–2001.
Marcoulides, K. M. (2017). *A Bayesian synthesis approach to data fusion using augmented data-dependent priors* (Unpublished doctoral dissertation). Arizona State University.
Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis, 10,* 292–304.
Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association, 83,* 1023–1032.

NCES. (2018). *Early Childhood Longitudinal Program (ECLS)*—Overview. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education. Retrieved from https://nces.ed.gov/ecls/
OECD. (2002). *PISA 2000 technical report*. Paris: Organization for Economic Cooperation and Development.
OECD. (2019). PISA 2018 Results: (Volumes I-IV): What students know and can do. https://doi.org/10.1787/5f07c754-en
Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics, 11,* 5018–5051. https://doi.org/10.1214/17-EJS1337SI.
Pindyck, R. S., & Rubinfeld, D. L. (1991). *Econometric models & economic forecasts*. McGraw-Hill.
Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases, 29,* 175–188.
R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics, 70*(4), 1023–1032.
Stan Development Team. (2021). *RStan: the R interface to Stan*. Retrieved from http://mc-stan.org/ (R package version 2.21.3)
Theil, H. (1966). *Applied economic forecasting*. Noth-Holland.
Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K), combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks (NCES 2009-004)*. U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics.
Trendtel, M., & Robitzsch, A. (2021). A Bayesian item response model for examining item position effects in complex survey data. *Journal of Educational and Behavioral Statistics, 46*(1), 34–57.
Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. Retrieved from https://CRAN.R-project.org/package=loo (R package version 2.1.0)
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27,* 1413–1432.
Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., & Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics, 13,* 41–54.
Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *TEST, 5,* 1–60.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com