On the Development of Probabilistic Projections of Country-level Progress to the UN SDG Indicator of Minimum Proficiency in Reading and Mathematics

David Kaplan[1*], Nina Jude[2], Kjorte Harra[1], Jonas Stampka[2]

[1*]Department of Educational Psychology, University of Wisconsin - Madison, 1025 W. Johnson St., Madison, 53706, Wisconsin, United States.

[2]Institute for Bildungswissenshaft, Universität Heidelberg, Akademiestr. 3, Heidelberg, 69117, Baden-Württemberg, Germany.

*Corresponding author(s). E-mail(s): david.kaplan@wisc.edu;
Contributing authors: jude@ibw.uni-heidelberg.de; harra@wisc.edu;
stampka@ibw.uni-heidelberg.de;

Abstract

As of this writing, there are five years remaining for countries to reach their Sustainable Development Goals deadline of 2030 as agreed to by the member countries of the United Nations. Countries are, therefore, naturally interested in projections of progress toward these goals. A variety of statistical measures have been used to report on country-level progress toward the goals, but they have not utilized methodologies explicitly designed to obtain optimally predictive measures of *rate of progress* as the foundation for projecting trends. The focus of this paper is to provide Bayesian probabilistic projections of progress to SDG indicator 4.1.1, attaining minimum proficiency in reading and mathematics, with particular emphasis on competencies among lower secondary school children. Using data from the OECD PISA, as well as indicators drawn from the World Bank, the OECD, UNDP, and UNESCO, we employ a novel combination of Bayesian latent growth curve modeling Bayesian model averaging to obtain optimal estimates of the rate of progress in minimum proficiency percentages and then use those estimate to develop probabilistic projections into the future overall for all countries in the analysis. Four case study countries are also presented to show how the methods can be used for individual country projections.

Keywords: UN Sustainable Development Goals, Model Uncertainty, Probabilistic Projections

1 Introduction

In 2000, member states of the United Nations adopted the *Millennium Development Goals* (MDGs) (United Nations, 2015). Eight goals were advanced, and of relevance to this paper, Goal 2 focused on achieving *universal primary education*. Previously, the 2000 World Education Forum in Dakar, Senegal set out the ambitious goal of [e]liminating gender disparities in primary and secondary education by 2005 and achieving gender equality in education by 2015, with a focus on ensuring that girls enjoy full and equal access to basic education of good quality (UNESCO, 2015).

Under Goal 2 of the MDGs, the specific target (Target 2a) specified that countries "(e)nsure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling.

By the end of 2015, considerable progress had been made. According to United Nations (2015), substantial progress was realized in (a) increasing enrollment rates, particularly in sub-Saharan Africa; (b) reducing the number of out-of-school children of primary school age; (c) increasing literacy rates world-wide, along with a reduction in the gender gap in reading literacy. Nevertheless, in the developing regions of the world, children in the poorest households were still four times as likely to be out of school as those in the richest households.

As the Millennium Development Goals came and went, the UN set about to again stimulate member countries to embrace and achieve the so-called *Sustainable Development Goals (SDGs)*. Again, with a focus on education, UNESCO together with UNICEF, the World Bank, UNFPA, UNDP, UN Women and UNHCR convened the World Education Forum 2015 in Incheon, Republic of Korea. Over 1,600 participants from 160 countries, including over 120 ministers, heads and members of delegations, heads of agencies and officials of multilateral and bilateral organizations, and representatives of civil society, the teaching profession, youth and the private sector, adopted the Incheon Declaration for Education 2030, which, at the time, set out a new vision for education for the next fifteen years (UNESCO, 2016a). Specifically, the UN identified equitable, high-quality education, including the achievement of literacy and numeracy by all youth and a substantial proportion of adults, both men and women, as one of its global SDGs to be attained by 2030 UN General Assembly (2015). As of this writing, there are five years remaining to achieve the SDGs generally and the education goals specifically.

2 SDG 4: Quality Education

The Incheon Declaration recognized that education is the primary driver in meeting many, if not all, of the other SDGs. The focus of this paper is SDG 4, indicator 4.1.1 which stipulates that the "[p]roportion of children and young people (a) in grades 2 & 3; (b) at the end of primary; and (c) at the end of lower secondary [achieve] at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex", and specifically 4.1.1c which concerns the end of lower secondary school.

To analyze the potential impact of education policies targeting the SDGs, it is critically important to monitor trends in educational outcomes over time. Indeed, as educational systems around the world face new challenges due to the lingering impact of the COVID-19 pandemic, monitoring trends in educational outcomes could help identify the long-run impact of this unprecedented health crisis on global education. To this end, international large-scale assessment (ILSA) programs such as the Organization for Economic Cooperation and Development (OECD) Program for International Student Assessment (PISA, OECD, 2001 are uniquely situated to provide population-level trend data on literacy and numeracy outcomes (Montoya & MacDonald, 2022).

2.1 Purpose and Significance

The purpose of this paper is to advance an approach for developing probabilistic projections of the rate of progress that countries are making in reaching or exceeding *minimum proficiency levels* in reading and mathematics.[1] We rely on data from PISA which assesses in-school 15-year-old students, and hence provides evidence of relevance to SDG indicator 4.1.1c. Following a "proof-of-concept" paper by Kaplan & Huang (2021), we will first derive an estimate of the rate of progress in minimum proficiency across countries using Bayesian latent growth curve modeling. We will then draw on country-level indicators from PISA as well as data sources from UNESCO and the World Bank to create an initial set of predictors of the rate of progress for boys and girls separately within and across countries. We will then apply *Bayesian*

[1] From here on, we will simply use the term minimum proficiency rather than minimum proficiency levels.

model averaging (BMA), described below, to account for model uncertainty in the selection of the choice of indicators of rate of progress. We will then use Bayesian model averaged predictors of the rate of progress to generate optimal probabilistic predictions of minimum proficiency in reading and mathematics as defined in PISA.

The significance of this paper is three-fold. First, we view the methodologies applied in this paper as contributing to monitoring progress in the internationally agreed-upon education goals of UNESCO, the OECD, the World Bank, the World Education Forum, and others. Second, we view this paper as contributing to sophisticated secondary analyses of large-scale educational assessments. This paper demonstrates the richness of policy information that can be obtained when using Bayesian probabilistic prediction models to study educational trends at the population level. Third, this paper is designed to advance Bayesian methodology in education through a novel synthesis of longitudinal models for trend projections with methods that account for uncertainty in said projection models.

2.2 Organization

The organization of this paper is as follows. In Section 3 we provide an overview of current international trends in reading and mathematics proficiencies along with trend plots for the countries that will be used in this study. Section 4 introduces the method of Bayesian latent growth curve modeling that we will use for estimating rate of progress in achieving minimum proficiency. In Section 5, we provide a brief review of Bayesian decision theory which motivates the method we use to obtain optimal predictions of growth. In Section 6 we introduce BMA which is the method we use to obtain optimal estimates of growth while taking into account model uncertainty. Section 7 discusses the data sources, selection of indicators, and methods for handling missing data, along with the workflow for this paper. We introduce the main data source used to assess change over time in the country-level percentages of students achieving minimum proficiency in reading and mathematics, namely PISA[2]. Section 8 presents the growth curve modeling and the model averaging results. Section 9 presents the overall projections across the countries used in this analysis as well as our case studies. Section 10 concludes.

3 Overview of Trends in Reading and Mathematics Literacy

A major substantive focus of this paper is educational equity, and in particular the impact of COVID-19 on exacerbating inequities in educational outcomes. Educational equity at the system level can be defined as schools providing equal learning opportunities for all students (OECD, 2018). Equity can be seen as a fundamental goal of education and a guiding principle in education policy that is accepted worldwide (UNESCO, 2018). In an ideal world, students' characteristics, such as family background or gender, should be unrelated to educational outcomes such as literacy in different domains. However, countries clearly differ in their level of equity already attained as well as in the national policies they implement to reach agreed-upon goals.

Policy measures promoting equity can be directed at the individual, the institution, or the system level (Levin, 2003). Educational systems that are less stratified and provide more flexible educational pathways tend to show more equity in educational outcomes (OECD, 2008). Opposite effects can be seen in systems where parents (or students) are given free school choice. This often leads to increased differences in the social composition of schools, which is usually related to aspects of equity. For example, gender gaps can be lower in high-socioeconomic status schools (Borgonovi et al., 2018).

In the case of gender equity as a specific goal, significant progress has been made and the global impact of this progress can be seen in current education statistics. For

[2]Unless otherwise stated, for simplicity we will use the phrase "achieving minimum proficiency" where it is understood that we are referring to both boys and girls and for both reading and mathematics

example, with regard to school enrollment rates, parity was achieved globally, on average, up to upper-secondary education (UNESCO, 2016b). Still, gender gaps in literacy exist in all countries where data are available; although girls usually outperform boys in reading, the opposite is the case for mathematics and science. While the actual gender gap in students' literacy could be considered comparatively small in many of the OECD countries, these effects do not vanish, but rather increase over time. For example, results from the OECD Program for the International Assessment of Adult Competencies (PIAAC), which focuses on adult populations, showed that competence in numeracy was significantly higher for men than women in all participating countries (Encinas-Martín & Cherian, 2023).

While graduation rates in tertiary education are now comparable between men and women in most OECD countries (Schleicher, 2008), gender differences related to the STEM subjects remain high. The result is a disadvantage for women in accessing this rapidly developing share of the labor market (UNESCO, 2016b).

We have so far described the state-of-affairs prior to the COVID-19 pandemic. This unprecedented health crisis has the potential of exacerbating the gender disparities in education between men and women. Current research already points to significant impacts of the pandemic on educational outcomes that will probably be seen in the years to come (NCES, 2022; World Bank et al., 2022). Connected to the impacts of the pandemic on educational outcomes, data collected by the United Nations shows that the impact of the pandemic threatens to deepen gender poverty gaps (UNDP, 2020). However, effects might be country specific, as the TIMSS 2023 results show that, in many countries, the performance in mathematics and science remained stable or improved compared to 2019, while others experienced learning losses particularly in mathematics (von Davier et al., 2024).

Clearly, challenges remain in identifying the origins of gender differences in major literacy domains. Policymakers need to be aware of the long-term influence of gender gaps existing at school-age (Borgonovi et al., 2017), and especially now in tracking gender gaps resulting from the pandemic. Consequently, tracking the development of gender equity over time is crucial to international educational policy.

We will focus on country-level longitudinal outcomes in minimum proficiency using data from PISA 2009 to PISA 2022. Although longer time points are available, it was decided to only include countries with complete data on the reading and mathematics outcomes so that predictive analyses were not dependent on the imputation of excessive amounts of missing data. The list of these countries is shown in Table 1.

Table 1: Fifty-three countries and economies with complete reading and mathematics assessment data from 2009 to 2022.

Albania	Argentina	Australia	Austria	Belgium
Brazil	Bulgaria	Canada	Chile	Colombia
Croatia	Czech Republic	Denmark	Estonia	Finland
France	Germany	Greece	Hong Kong–China	Hungary
Iceland	Indonesia	Ireland	Israel	Italy
Japan	Jordan	Kazakhstan	Korea	Latvia
Lithuania	Macao–China	Mexico	Montenegro	Netherlands
New Zealand	Norway	Peru	Poland	Portugal
Qatar	Romania	Singapore	Slovak Republic	Slovenia
Spain	Sweden	Switzerland	Thailand	Turkey
United Kingdom	United States	Uruguay		

Note: Luxembourg did not have data for PISA 2022, and the Russian Federation was excluded from the PISA 2022 assessment due to the war in Ukraine.

Two points are worth noting regarding the countries shown in Table 1. First, these are countries who are either members of the OECD or partner countries that are participating in PISA, and although there is a range of economic development covered by these countries, all of these countries are considered economic democracies as defined by the OECD. Second, it is important to note that much of the global south countries are not represented in this list; countries in sub-Saharan

Africa, the Indian sub-continent, and China. This is relevant insofar as the vast majority of the world's children are located in China and India. The major reason for the exclusion of these countries is simply the lack of data, and where there are data sources at regional levels, the ability to provide reliable links to PISA remains an open methodological question. This issue will be taken up in the Conclusions section.

The longitudinal outcome will be the percentage of students at the country level, male and female, who have reached the PISA-defined minimum proficiency level across five cycles of PISA. For reading, the PISA minimum proficiency lower cut score is 407 defining *Level 2* of the PISA proficiency scale. For mathematics, the lower cut score for minimum proficiency is 420. Figures (1) and (2) show the changes over time in these percentages.

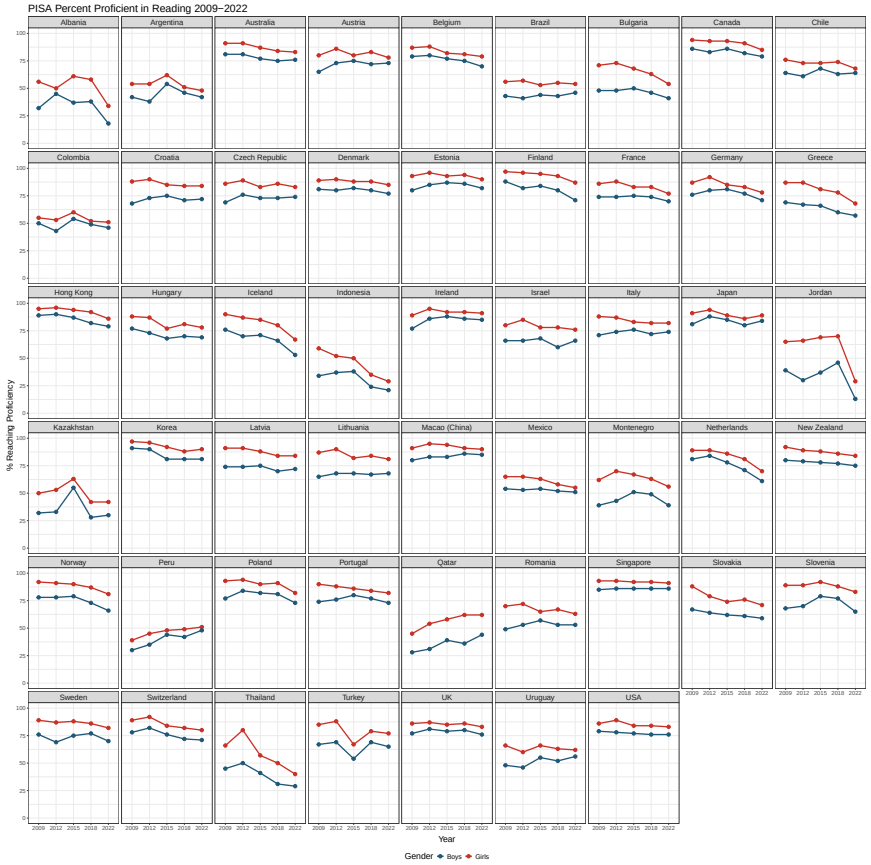

Fig. 1: Trend lines for percentage of students at the minimum proficiency level in reading from 2009 to 2022. The red line is for the girls, the blue line is for boys. A downward slope indicates a decrease in the percentage attaining the minimum proficiency level in reading.

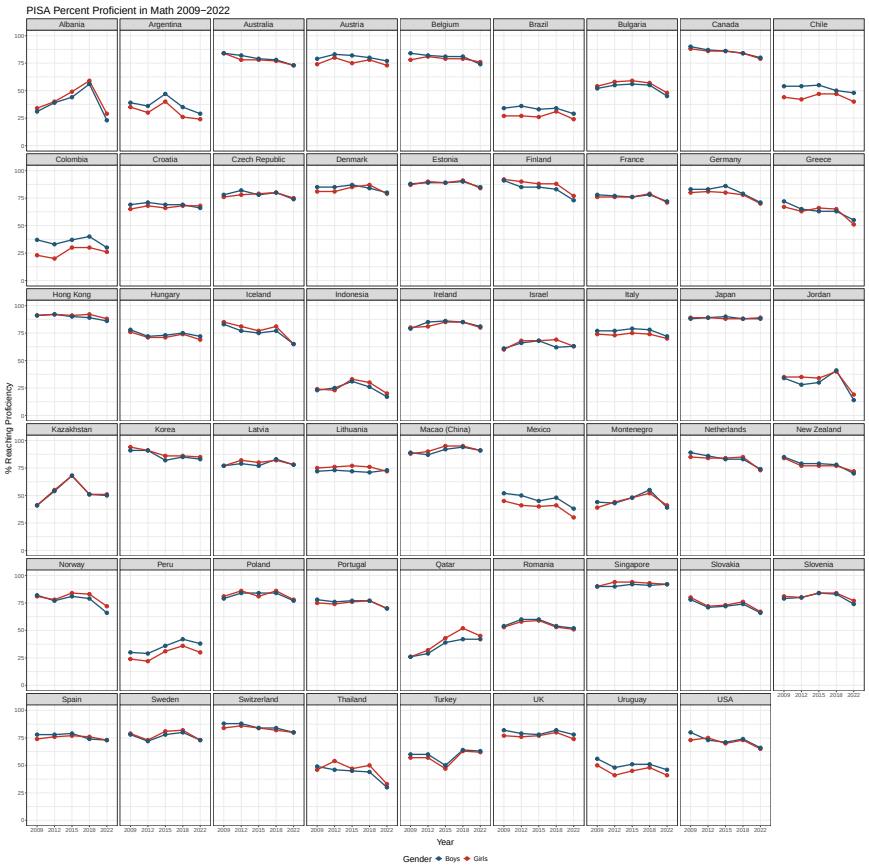

Fig. 2: Trend lines for percentage of students at the minimum proficiency level in mathematics from 2009 to 2022. The red line is for the girls, the blue line is for boys. A downward slope indicates a decrease in the percentage attaining the minimum proficiency level in mathematics.

Inspection of Figures (1) and (2) reveal some interesting features. Figure (1) shows the well-documented differences between the percentage of girls and boys at the minimum proficiency in reading. That said, the trend over time for boys and girls, while somewhat variable across countries, are basically parallel to each other within countries with some countries showing gains in the percentage at minimum proficiency and others declining. For mathematics proficiency shown in Figure (2), the advantage favoring boys is not as clear, and it appears that again, although there is some variability in the trend across countries, very little difference is observed between boys and girls over years 2009 to 2022. Finally, for both reading and mathematics, and for both boys and girls, the shape of the trend lines is mostly linear for the countries in this analysis with a general downward trend in most countries. Some countries exhibit a steep decline in 2022, which could be due to the impact of the pandemic, but definite causal conclusions cannot be drawn.

4 Bayesian Growth Curve Modeling

With the empirical trajectories in hand, the next step is to estimate the average starting level and rate of progress in the percent at or above minimum proficiency over the cycles of PISA. We frame our estimation of these parameters as a Bayesian hierarchical latent variable model (Kaplan, 2023).

Let the *within-country* model be written as follows

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\,\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_{y(i)} \tag{1}$$

where $\boldsymbol{y}_i$ is a $T \times 1$ vector of T waves of measurement for country i ($i = 1, \ldots, n$); $\boldsymbol{\Lambda}$ be a $T \times Q$ matrix of fixed constants that for simplicity are assumed to be the same for all countries. These terms can be specified to represent linear, quadratic, or even higher order shapes of the trajectories and also serve to parameterize the growth model as a structural equation model (Willett & Sayer, 1994; Muthén, 1997; Bollen & Curran, 2006; Grimm et al., 2017). Further, let $\boldsymbol{\eta}_i$ be a $Q \times 1$ vector of random growth parameters for country i, where the random growth parameters can be considered as latent variables. In our study, $Q = 2$, representing the intercept and linear growth component for each country; and $\boldsymbol{\epsilon}_{y(i)}$ is a $T \times 1$ vector of residuals, with diagonal covariance matrix

$$\boldsymbol{\Sigma}_{y_i} = \begin{bmatrix} \sigma^2_{y_{1(i)}} & & & & \\ & \sigma^2_{y_{2(i)}} & & & \\ & & \ddots & & \\ & & & \sigma^2_{y_{(T-1)(i)}} & \\ & & & & \sigma^2_{y_{T(i)}} \end{bmatrix}. \tag{2}$$

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

$$\boldsymbol{\eta}_i = \boldsymbol{\Gamma}\boldsymbol{x}_i + \boldsymbol{\epsilon}_{\eta(i)}, \tag{3}$$

where $\boldsymbol{\Gamma}$ is a $Q \times P$ vector of regression coefficients associated with the time-invariant predictors, $\boldsymbol{x}_i$ is a $P \times 1$ vector of time-invariant predictors, and $\boldsymbol{\epsilon}_{\eta(i)}$ is a $Q \times 1$ vector of residuals with symmetric covariance matrix

$$\boldsymbol{\Sigma}_{\eta_i} = \begin{bmatrix} \sigma^2_{\eta_0} & sym. \\ \sigma_{\eta_1\eta_0} & \sigma^2_{\eta_1} \end{bmatrix}, \tag{4}$$

allowing for the random growth parameters (i.e., intercept and slope) to be correlated conditional on the predictors as show in Eq. (4).

4.1 Bayesian Hierarchical Growth Curve Model

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

$$\boldsymbol{y}_i \sim N(\boldsymbol{\Lambda}\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_{y(i)}), \tag{5a}$$
$$\boldsymbol{\eta}_i \sim N(\boldsymbol{\Gamma}\boldsymbol{x}_i, \boldsymbol{\Sigma}_{\eta_i}), \tag{5b}$$
$$\boldsymbol{\Gamma} \sim N(\boldsymbol{\Omega}, \boldsymbol{\Sigma}_{\Gamma}), \tag{5c}$$

where $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}_{\Gamma}$ are fixed and known parameters. Prior distributions on the residual covariance matrices are assumed to be inverse-Wishart (IW). For the current data, let $\sigma^0_{y_i}$ indicate the standard deviation of level-1 residual, which is the square root of the diagonal components in Eq. (2). The prior distributions for the variance parameters are specified as follows:

$$\sigma^0_{y_i} \sim \text{half-Cauchy}(\mu_y, \sigma_y), \tag{6a}$$
$$\boldsymbol{\Sigma}_{\eta_i} \sim \text{IW}(\mathbf{R}_\eta, \nu_\eta), \tag{6b}$$

but of course other prior specifications are possible (see Kaplan, 2023).

4.2 Flexible Models for rate of progress

An important flexibility in growth curve modeling allows for the estimation of non-linear trajectories using so-called *latent basis* methods. This specification requires

that some of the time points be fixed to constant values for model identification purposes while allowing the remaining time points to be estimated from the data directly. Latent basis modeling yields data-based estimates of the time points and often provides better fit of the model to the empirical trajectories of change over time compared to forcing, say, a linear trend on the data. Within the framework of latent basis methods, the rate parameter π_{1i} is best conceived of as a *shape* parameter, but we will continue to refer to this as a rate of progress parameter. These, and other extensions, are discussed in Bollen & Curran (2006) and Grimm et al. (2017).

For this paper, we will examine differences in predictive performance across three latent growth curve models: (a) a simple linear growth curve model with fixed and equally spaced time points, (b) a latent basis model with that last time point (2022) freed to reflect a possible decrease in scores due to the COVID-19 pandemic, and (c) a latent basis model in which the basis functions from 2015-2022 are freed, reflecting the idea that the downward trend had started prior to 2022 (see also Kaplan et al., 2025)

For this paper, the inter-country model for the starting minimum proficiency percentage π_{0i} and rate of progress π_{1i} respectively, can be written as

$$\pi_{0i} = \beta_{00} + \sum_{q=1}^{Q} \beta_{0q} x_{qi} + \epsilon_{\pi_{0i}}, \qquad (i = 1, \ldots, n) \tag{7}$$

and

$$\pi_{1i} = \beta_{10} + \sum_{q=1}^{Q} \beta_{1q} x_{qi} + \epsilon_{\pi_{1i}}, \qquad (i = 1, \ldots, n) \tag{8}$$

where x_{qi} are values on Q predictors for country i, β_{00}, β_{01}, β_{10}, and β_{1q} are the intercept and slopes associated with the time-invariant predictors of the starting minimum proficiency percentage and rate of progress, and $\epsilon_{\pi_{0i}}$ and $\epsilon_{\pi_{1i}}$ are errors. An example of a time-invariant predictor might be the type of political system of a country.[3] For this paper, we will develop a set of distinct models for predicting π_0 and π_1.[4]

5 The Decision-Theoretic Framework for Predictive Modeling

Foundational to our goal of developing an *optimally predictive* model to be used to forecast the rate of progress in minimum proficiency is *Bayesian decision theory*. Bayesian decision theory structures the problem of predictive modeling in the context of minimizing *expected loss* – that is, the penalty that results from using a particular model to predict future observations. The less the expected loss, the better the model is at predictive performance in comparison to other models. Bayesian decision theory (see e.g., Good (1952); Lindley (1991); Berger (2013)) provides a natural and intuitive foundation for developing and evaluating Bayesian predictive models.

5.1 Fixing Notation and Concepts

An outline of our Bayesian decision-theoretic framework was given in Kaplan et al. (2025) and is reproduced here. Let $D = \{y_{ci}, x_i, z_{ci}\}_{i=1}^{n}$ be a set of data that is assumed to be fixed in the Bayesian sense where one conditions on observable and fixed data, and where y_{ti} is the outcome of interest at cycle t for country i, x_i is a (possibly vector-valued) set of time-invariant predictors, and z_{ti} is a set of time varying-predictors for country i. Additionally, let $(\tilde{y}, \tilde{x}, \tilde{z})$ be a future observation of the outcome of interest and the set of predictors, respectively. Finally, let $\mathcal{M} = \{M_k\}_{k=1}^{K}$ represent a set of individual models that are specified to yield predictions of the rate of progress in minimum proficiency, π_1 and let M_k represent a

[3] Of course, this assumes that the political system of a country is relatively stable over time.

[4] Note that Equations (7) and (8) imply that the same predictors are being used for the starting minimum proficiency percentage and rate of progress, and although that will be the case for this paper, it is not necessary, and different predictors for these parameters can be specified.

specific model for the rate of progress. Each M_k will eventually be a member in the ensemble $\mathcal{M}$.

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

To provide a context for these ideas, and in anticipation of our application to minimum proficiency for indicator 4.1.1c, consider the problem of predicting a future percentage of boys and girls reaching or exceeding minimum proficiency. Following Bernardo & Smith (2000), Lindley (1991), Vehtari & Ojanen (2012) and Berger (2013) and the notation given previously for country i at time t, (a) the states of the world correspond to the future minimum proficiency percentages from future cycles of PISA, that is, $\tilde{y} \in \mathcal{Y}$, (b) the action $a \in \mathcal{A}$ is the actual prediction of those future observations based on using an optimized prediction of the rate of progress π_1, (c) the loss function $L(a, \tilde{y})$ defines the loss attached to the prediction, and (d) a posterior predictive distribution, $p(\tilde{y}|D, M_*)$, that encodes our belief about the rate of progress in minimum proficiency percentages conditional on the data, D.

5.2 Loss Functions for Evaluating Predictions

The intent of predictive modeling is to minimize the loss associated with taking an action a among a set of actions included in an action space $\mathcal{A}$. A number of loss functions exist, but common loss functions rely on the negative *quadratic loss* function

$$L(a, \tilde{y}) = (\tilde{y} - a)^2. \tag{9}$$

The optimal action a^* is the one that minimizes the *posterior expected loss*, written as

$$a^* = \arg\min_{a \in \mathcal{A}} \int_{\Omega} L(\omega, a) p(\omega|D) d\omega. \tag{10}$$

The objective of predictive modeling, therefore, is to take an action a that minimizes the loss L when the future observation is $\tilde{y}$. Clyde & Iversen (2013) showed that the optimal decision a^* obtains when $a^* = E(\tilde{y}|D)$, which is the posterior predictive mean of $\tilde{y}$ given the data D. Assuming that the correct data generating model exists and is among the set of models under consideration, this can be expressed as

$$E(\tilde{y}|D) = \sum_{k=1}^{K} E(\tilde{y}|M_k, D) p(M_k|D) = \sum_{k=1}^{K} p(M_k|D) \hat{\tilde{y}}_{M_k}, \tag{11}$$

where $\hat{\tilde{y}}_{M_k}$ is the posterior predictive mean of $\tilde{y}$ under M_k and $p(M_k|D)$ is the *posterior model probability* (PMP) associated with model k. The PMP can be written as

$$p(M_k|D) = \frac{p(D|M_k) p(M_k)}{\sum_{l=1}^{K} p(D|M_l) p(M_l)}, \qquad l \neq k \tag{12}$$

where $p(M_k)$ the prior probability for model k. Equations (11) - (12) define BMA.[5] By default, a uniform prior probability mass is placed over the model space and non-informative unit information priors (Kass & Wasserman, 1996) are used for the parameters of each model, but software programs such as BMS (Zeugner &

[5] The assumption that the true model is in the space of models under investigation is referred to as the $\mathcal{M}$-closed framework (Bernardo & Smith, 2000). This a rather difficult assumption to warrant in the education and the social sciences generally. A relaxation of this assumption is referred to as the $\mathcal{M}$-open setting (Bernardo & Smith, 2000). A recent paper with applications to PISA 2018 by Kaplan (2021) compared BMA with *Bayesian stacking* (e.g.Breiman (1996), which is suited for the $\mathcal{M}$-open settings and found BMA to provide slightly better predictive performance. Thus, in following previous studies that have applied BMA to to growth regression in economics (e.g. Fernández et al. (2001b) we will use BMA for this study.

Feldkircher, 2015) allow for other prior distribution choices over both the parameters and the model space. We will examine the sensitivity of our BMA analysis to the choice of priors in our empirical analysis below.

6 Bayesian Model Averaging

When addressing the problem of developing probabilistic projection models, it is common to specify and estimate a set of different projection models and to use various model selection methods such as Akaike's information criterion (AIC) (Akaike, 1985, 1987) or the Bayesian information criterion (BIC) (Schwarz, 1978; Kass & Raftery, 1995) to choose a final model to report. The difficulty with model selection methods is that the analyst often proceeds as though the final selected projection model was the one considered in advance, and thus the uncertainty in the model selection process is ignored. More to the point, the selection of a particular model from a universe of possible models can be characterized as a problem of uncertainty. The problem of model uncertainty has been succinctly characterized by Hoeting and her colleagues (Hoeting et al., 1999):

> Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are. (p. 382)

To address the problem of model uncertainty, we propose to use BMA (Leamer, 1978; Madigan & Raftery, 1994; Raftery et al., 1997; Hoeting et al., 1999). The key idea of BMA (described in more detail below) is to recognize that selecting a single model out of a class of possible models that could have been selected effectively ignores the uncertainty inherent in model choice. Model averaging diminishes the uncertainty in the choice by taking a weighted average over a selected set of models, with weights that account for the quality of the model. These weights are the posterior model probabilities (PMPs).

Of importance to this paper is that theory and applications of BMA have shown that it provides better predictive performance to that of any single model based on a class of scoring rules used in probabilistic forecasting (Raftery et al., 1997). Below, we will employ two specific scoring rules that we will use to evaluate the predictive performance of our projection models.

Bayesian model averaging has had a long history of theoretical developments and practical applications. Early work by Leamer (1978) laid the foundation for BMA. Fundamental theoretical work on BMA was conducted in the mid-1990s by Madigan and his colleagues (e.g., Madigan & Raftery (1994); Raftery et al. (1997); Hoeting et al. (1999)). Additional theoretical work was conducted by Clyde (1999). Draper (1995) has discussed how model uncertainty can arise even in experimental designs, and Kass & Raftery (1995) provide a review of BMA and the consequences of ignoring model uncertainty. A more general review of the problem of model uncertainty can be found in Clyde & George (2004).

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

6.1 Technical Background of BMA

The aim of this paper is to provide a set of methods and practices that can yield probabilistic projections of country-level progress in meeting or exceeding minimum proficiency in reading and mathematics. Following Madigan & Raftery (1994), we denote a future outcome of interest (in our case, country-level mathematics or reading minimum proficiency) as Υ. Next, consider a set of competing models $\{M_k\}_{k=1}^{K}$ that are not necessarily nested. The posterior distribution of Υ given data y can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y), \tag{13}$$

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

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l \neq k. \tag{14}$$

The interesting feature of Equation (14) is that $p(M_k|y)$ will likely be different for different models. The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{15}$$

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

6.2 Computational considerations for BMA

The software program we will use in this paper is BMS (Zeugner & Feldkircher, 2015), which is part of the R programming environment (R Core Team, 2023). The default algorithm within BMS is referred to as the birth-death (BD algorithm). The BD algorithm a Metropolis-Hastings-type Markov Chain Monte Carlo (MCMC) method designed to explore the vast space of possible regression models. In BMA, each model corresponds to a specific subset of covariates. Since the total number of models grows exponentially with the number of covariates (e.g., 2^K for K predictors), the BD algorithm samples models rather than enumerating all of them.[6] It does this by proposing local moves: a *birth* move adds one variable to the current model, while a *death* removes one variable. These moves ensure that the Markov chain eventually approximates the posterior distribution over the model space.

At each step, the algorithm randomly chooses whether to attempt a birth or a death move (with roughly equal probability). It then proposes a new model by either adding a new covariate (birth) or removing one (death), and calculates the posterior model probability (PMP) of the proposed model. The move is accepted with a probability that depends on the ratio of the posterior probabilities of the new and current models, adjusted for the proposal probabilities. This acceptance mechanism ensures that models with higher posterior probabilities are more frequently visited, allowing the empirical frequency of visits to approximate the true posterior model probabilities.

The BD algorithm is particularly efficient in large model spaces because it focuses sampling on high-probability regions while still allowing occasional exploration of lower-probability models. Over many iterations, this results in a posterior distribution over models and predictors that can be summarized in terms of posterior inclusion probabilities (PIPs), predictive densities, and model rankings. The algorithm naturally favors models with strong support in the data, and if the true data generating model is in the space of models under consideration, then the BD algorithm will asymptotically concentrate its sampling around the true model, if

[6] This does not count interactions in which case the model space is $2^{K+\binom{K}{2}}$. Interactions can be handled in the BMS program but we omit them in this study.

the true model is not in the space, on models with highest posterior mass (see also Footnote 4).

6.3 Choosing Parameter and Model Priors in BMA

A required step when implementing BMA is to choose both parameter and model priors (Fernández et al., 2001a; Liang et al., 2008; Eicher et al., 2011; Feldkircher & Zeugner, 2009). In BMA, parameter priors are variations of Zellner's g-prior (Zellner, 1986). Specifically, Zellner introduced a natural-conjugate Normal-Gamma g-prior for regression coefficients β under the normal linear regression for model, written as,

$$y_i = x_i'\beta + \varepsilon, \tag{16}$$

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

$$\beta_k|\sigma^2, M_k, g \sim N\left(0, \sigma^2 g\left(x_k'x_k\right)^{-1}\right). \tag{17}$$

Implementing the g-prior requires a choice between *fixed* forms of the g-prior and *flexible* forms of the g-prior Feldkircher & Zeugner (2009). Fixed forms of the g-prior require that the g be set to a particular value, typically related to the sample size, the number of predictors or both. These include (a) the *Unit Information Prior* (UIP, $g = N$); (b) the *Risk Inflation Criterion Prior* (RIC, $g = Q^2$); (c) the *Benchmark risk inflation criterion* (BRIC, $g = \max(N, Q^2)$; and (d) the *Hannan and Quinn Priors* (HQ, $g = \log(N)^3$).

As pointed out by Feldkircher & Zeugner (2009), a problem with fixed priors is that they are *dengenerate* in the sense that all the prior sets all the probability to one point. Instead of only employing degenerate priors, Liang et al. (2008) introduced the *hyper-g-prior* that "let the data choose" (see also Feldkircher & Zeugner, 2009), thus reducing the sensitivity of the prior choice of the fixed g-prior on the posterior mass. This family of priors was proposed for data-dependent shrinkage. Following Feldkircher & Zeugner (2009), the hyper-g prior is a Beta prior on the shrinkage factor $\frac{g}{1+g}$, that is $p\left(\frac{g}{1+g}\right) \sim Beta(1, \frac{\alpha}{2} - 1)$, with $E\left(\frac{g}{1+g}\right) = \frac{2}{\alpha}$. Instead of eliciting g directly,the hyper-g prior requires the elicitation of the hyperparameter $\alpha \in (2, \infty)$. As α approaches 2, the prior distribution on the shrinkage factor $\frac{g}{1+g}$ will be close to 1; while for $\alpha = 4$, the prior distribution on the shrinkage factor will be uniform distributed. In the context of noisy data, the hyper-g prior will distribute posterior model probabilities more uniformly across the model space. In the case of low noise in the data, the hyper-g prior will be concentrated on a few models, and perhaps even more concentrated than in the fixed prior case with large g (Feldkircher & Zeugner, 2009).

For this paper, three model priors are investigated and are available in the BMS program: (a) the uniform model prior, (b) the binomial model prior, and (c) the beta-binomial model prior. The uniform model prior is a common default prior for Bayesian model averaging. Specifically, if there are Q predictors, then the prior on the space of models is 2^{-Q}. The difficulty with the uniform model prior was pointed out by Zeugner & Feldkircher (2015) who noted that the uniform model prior implies that the expected model size is $\sum_{q=0}^{Q} \binom{Q}{q} q 2^{-Q} = Q/2$. However, the distribution of model sizes is not even. The result is that the uniform model prior actually places more mass on intermediate size models. A demonstration of the impact of this problem is given in Zeugner & Feldkircher (2015).

To address the problem with the uniform model prior, Zeugner & Feldkircher (2015) proposed placing a fixed inclusion probability on each predictor in the model, denoted as θ. Then, for model k, the prior probability for a model of size q is $p(M_k) = \theta^{q_k}(1-\theta)^{Q-q_m}$. Notice that the expected model size, say $\bar{m}$, is $Q\theta$, and thus the analysts prior expected model size is $\bar{m}$. Moreover, if $\theta = .5$, then the binomial model prior reduces to the uniform model prior. In practice, this suggests that choosing $\theta < .5$ weights the posterior mass toward smaller models, and visa versa Zeugner & Feldkircher (2015).

The binomial prior discussed above suffers from the fact that the inclusion probability θ is fixed. Following Ley & Steel (2009), greater flexibility is obtained by treating θ as random. A logical choice for the probability distribution of θ is the Beta distribution with hyperparameters $a, b > 0$, viz. $\theta \sim \text{Beta}(a, b)$, giving rise to the *beta-binomial* model prior

A number of studies have been conducted comparing the performance of combinations of parameter and model priors. It is beyond the scope of this paper to review each of these studies, however, the most recent, and arguably the most comprehensive studies to compare parameter and models priors were conducted by Porwal & Raftery (2022, 2023). With respect to parameter priors, Porwal & Raftery (2022) evaluated 21 widely used methods through extensive simulation studies based on real datasets representing diverse practical scenarios. Three adaptive Bayesian model averaging (BMA) methods emerged as the top performers across all statistical tasks. These methods utilized adaptive versions of Zellner's g-prior, where the prior variance parameter g is either estimated from the data or set as the sample size n. Two of the best-performing methods, BMA with $g = \sqrt{n}$ (see Fernández et al., 2001a) and local empirical Bayes, were found to be as efficient as LASSO (Tibshirani, 1996), a commonly preferred variable selection technique. Additionally, BMA outperformed Bayesian model selection, which selects only a single model.

With regard to model priors, Porwal & Raftery (2023) examined eight reference model space priors commonly used in the literature, along with three adaptive parameter priors. To evaluate their effectiveness in variable selection for linear regression models, Porwal & Raftery (2023) assessed their performance across key statistical criteria, including parameter estimation, interval estimation, inference, and both point and interval prediction. Their analysis was based on an extensive simulation study using 14 real datasets that captured a variety of practical scenarios. Porwal & Raftery (2023) found that beta-binomial model space priors, defined in terms of the prior probability of model size, performed best on average across different statistical criteria and datasets, surpassing priors that assign equal probability to all models. In contrast, recently introduced complexity priors (Castillo et al., 2015) showed relatively weaker performance.

For this paper, we examined BMA under a combination of parameter priors (UIP, RIC, BRIC, and HQ) and model space priors (uniform, binomial, and beta-binomial) using the Kullback-Leibler divergence (Kullback & Leibler, 1951; Kullback, 1959, 1987) as the measure of predictive performance (see Section 6.5). The results showed virtually no difference in terms of Kullback-Leibler divergence across all combinations of parameter and model space priors. Similar results were found in empirical applications by Kaplan & Huang (2021) and Kaplan (2021). Therefore, for the purposes of this paper, we utilize the UIP for parameter priors and the uniform distribution for the model space priors.

6.4 Scoring rules for BMA

A central characteristic of statistics is to develop accurate predictive models (Dawid, 1984). Indeed, as pointed out by (Bernardo & 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 element in the development of accurate predictive models is to decide on rules for gauging predictive accuracy – termed *scoring rules*. Scoring rules provide a measure of the accuracy of probabilistic projections, and a forecast can be said to be "well-calibrated" if the assigned probabilities of the outcome match the actual proportion of times that the outcome occurred.

A number of scoring rules are discussed in the literature (see e.g., Winkler (1996); Bernardo & Smith (2000); Jose et al. (2008); Merkle & Steyvers (2013); Gneiting & Raftery (2007)), however, for this paper we will primarily evaluate predictive performance under different parameter and model prior settings using the Kullback-Liebler Divergence (KLD) and the *leave-one-out information criterion*

(LOOIC). The KLD between two distributions can be written as

$$I(f,g) = \int f(x) log\left(\frac{f(x)}{g(x|\theta)}\right) dx, \tag{18}$$

where $I(f,g)$ is the information lost when g is used to approximate f. For this paper, the estimated growth rate without predictors is compared to predicted growth rate using BMA along with different choices of model and parameter priors. The projection model with the lowest KLD measure is deemed best in the sense that the information lost when approximating the "true" outcome distribution with the distribution predicted on the basis of the model is lowest. For this paper, LPS values will be obtained using BMS, and the KLD values will be obtained using the R package LaplacesDemon (Statisticat & LLC., 2021).

6.5 Leave-One-Out Cross-Validation

For this paper, we utilize *leave-one-out cross validation* (LOO-CV) to evaluate the predictive performance of the three growth curve models discussed earlier. Leave-one-out-cross-validation (LOO-CV) is a special case of q-fold cross-validation (q-fold CV) when $q = n$. In q-fold CV, a sample is split into q groups (folds) and each fold is taken to be the validation set with the remaining $q-1$ folds serving as the training set. For LOO-CV, each observation serves as the validation set with the remaining $n-1$ observations serving as the training set. Leave-one-out cross-validation is available in the R software program loo Vehtari et al. (2019).[7]

Following Vehtari et al. (2017) but contextualized in terms of our paper, let π_{1i} $(i = 1, \ldots, n)$ be an n-dimensional vector of slope parameters following a distribution conditional on parameters θ - viz. $p(\pi_1|\theta) = \prod_{i=1}^{n} p(\pi_{1i}|\theta)$. Given a prior distribution on the parameters, $p(\theta)$, we can obtain the posterior distribution, $p(\theta|\pi_1)$ as well as a posterior predictive distribution of predicted values of growth $\tilde{\pi}_1$ written as $p(\tilde{\pi}_1|\pi_1) = \int p(\tilde{\pi}_1|\theta)p(\theta|\pi_1)d\theta$. The Bayesian LOO-CV rests on the derivation of the *expected log point-wise predictive density* (ELPD) for new data defined as

$$\text{ELPD} = \sum_{i=1}^{n} \int p_t(\tilde{\pi}_{1i}) \log\, p(\tilde{\pi}_{1i}|\pi_1) d\tilde{\pi}_{1i}, \tag{19}$$

where $p_t(\tilde{\pi}_{1i})$ represents the distribution of the true but unknown data-generating process for each country's rate of progress toward minimum proficiency $\tilde{\pi}_{1i}$ and where Equation (19) is approximated by cross-validation procedures. The ELPD provides a measure of predictive accuracy for the n data points taken one at a time Vehtari et al. (2017). From here, the Bayesian LOO estimate can be written as

$$\text{ELPD}_{loo} = \sum_{i=1}^{n} \log p(\pi_{1i}|\pi_{1-i}), \tag{20}$$

where

$$p(\pi_{1i}|\pi_{1-i}) = \int p(\pi_1|\theta)p(\theta|\pi_{1-i})d\theta, \tag{21}$$

which is the leave-one-out predictive distribution using the log predictive score to assess predictive accuracy. An information criterion based on LOO (LOO-IC) can be derived as

$$\text{LOO-IC} = -2\,\widehat{\text{ELPD}}_{loo}. \tag{22}$$

which places the LOO-IC on the "deviance scale" (see Vehtari et al., 2017 for more details on the implementation of the LOO-IC in loo). Among a set of competing

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

models, the one with the smallest LOO-IC is considered best from an out-of-sample point-wise predictive point of view.

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

7 Data and Workflow

7.1 Indicator Selection

To select relevant indicators of countries' rate of progress towards the SDG targets, we extract data from several reliable sources. Following Lafortune et al. (2018), we prioritize indicators that are globally relevant, statistically valid, and up-to-date date to attempt to best capture current trends in progress towards the targets in reading and mathematics proficiency across countries. Our choice of indicators originate from the OECD, UNDP, UNESCO, and the World Bank which in turn derive from an idea in systems theory referred to as the *Context-Input-Process-Output* (CIPO) model of national educational systems (Scheerens, 2023). A conceptual diagram of the CIPO model is shown in Figure 3.

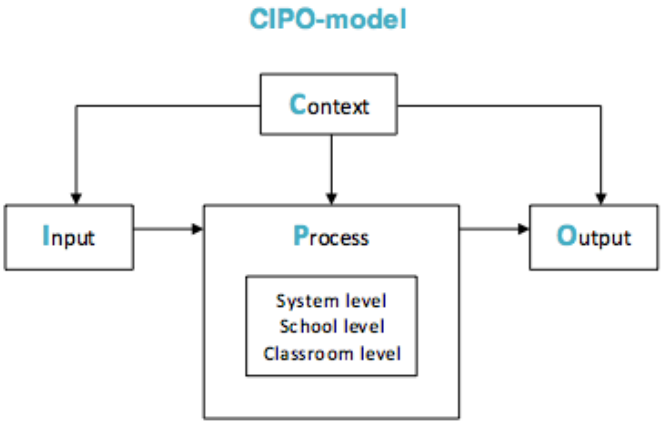

Fig. 3: Context-Input-Process-Output Model of Scherens (2000)

There are many ways to define quality education: Models include quality aspects regarding educational policies (OECD, 2022), school effectiveness (e.g. Scherens, 2000), teacher education, subject specific developments, and individual and social factors (Johnstone et al., 2020). Analysis of quality in education usually takes into account aspects of input, process, and outcomes, depending on the context of the school system (Scheerens et al. 2023) (see Figure 3). Hence, quality of education cannot be defined by outcomes alone, but must take into account contextual factors, input resources and processes which form the education system on different levels, leading to education outcomes and also can be seen as prerequisites for equity in education.

When analyzing the concrete SDG targets in relation to the CIPO model (see Table 2), it becomes clear that almost all of them either refer to outcome of education in the sense of the goals that are to be reached, or to aspects of access of education which can be seen as input factors or resources provided by an education system. Only target 4.c referring to a qualified teacher workforce could both be seen as an Input indicator but might also be related to the educational processes in schools under the assumption that qualification is needed to ensure quality in the

Table 2: List of indicators used for BMA categorized in terms of elements of the CIPO model.

Indicator	Member Models for Stacking	Difference Years
Employment females as percentage of employment annual	Context (Gender Gap)	2021 - 2009
Gender development index	Context (Gender Gap)	2022 - 2009
Gender gap index GEQ	Context (Gender Gap)	2022 - 2009
Labor force participation rate female percentage of female population	Context (Gender Gap)	2021 - 2009
Adolescents out of school of lower-secondary age	Outcome/Input (Education)	2022 - 2009
Children out of school of primary school age	Outcome/Input (Education)	2022 - 2009
Gross enrollment ratio primary both sexes	Outcome/Input (Education)	2022 - 2009
Gross enrollment ratio secondary both sexes	Outcome/Input (Education)	2022 - 2009
Lower-secondary school starting age years	Outcome/Input (Education)	2022 - 2009
Official entrance age to lower-secondary education years[1]	Outcome/Input (Education)	2022 - 2009
Official entrance age to pre-primary education years	Outcome/Input (Education)	2022 - 2009
Official entrance age to primary education years	Outcome/Input (Education)	2022 - 2009
GDP (standardized)	Context (SES)	2021 - 2009
Human development index	Context (SES)	2022 - 2009
Index highest occupational status of parents	Context (SES)	2022 - 2009
Rural population as percentage of total population	Context (SES)	2022 - 2009
Expected years of schooling	Context (Education)	2022 - 2009
Funding government	Context (Education)	2022 - 2009
Government expenditure on education percentage of GDP	Context (Education)	2022 - 2009
Government expenditure on education percentage of government expenditure	Context (Education)	2022 - 2009
Index school size	Context (Education)	2022 - 2009
Percentage of full-time teachers per school	Processes/Input (Education)	2022 - 2009
Percentage of part-time teachers per school	Processes/Input (Education)	2022 - 2009
Number of class periods in mathematics	Processes/Input (Education)	2022 - 2009
Primary school starting age years[1]	Processes/Input (Education)	2022 - 2009
Teaching hours lower secondary	Processes/Input (Education)	2021 - 2009
Teaching hours primary	Processes/Input (Education)	2021 - 2009
Teaching hours upper-secondary	Processes/Input (Education)	2021 - 2009
Percentage of teachers in pre-primary education who are female	Context (Education)	2022 - 2009
Percentage of teachers in primary education who are female	Context (Education)	2022 - 2009
Percentage of teachers in secondary education who are female	Context (Education)	2022 - 2009

[1] Denotes indicators removed from model averaging due to colinearity.

education setting. However, it can also be seen as in input factor as teacher qualification requires resources invested into teacher education itself.

Indicators relating to the context of education systems are not specifically stated under the SDG 4 targets. It can be assumed that context factors are rather country specific (like the structure of the school system) also might be changing rapidly (like flows of immigration or specifications for student admission). Moreover, indicators of context factors can also be found among the targets of other SDGs, for example indicators of poverty (SDG Goal 1).

Following the aspects of the CIPO model and linking it to the idea of predicting change in educational outcomes, respective indicators for input, context and processes of the education system would be needed for all countries reporting on their progress on the SDG. Hence, also when thinking about predicting progress towards the goals, indicators of input, context and processes need to be considered.

In addition to using the very few indicators already included in goal 4 of the SDG, the CIPO model can be used as a foundation to select additional indicators from other data sources which indicate aspects relevant to describe quality in education. These indicators of context, input, and processes can then be used to predict the change of the rate of progress towards the outcome indicators. Constructs being assessed in ILSA can provide the respective data.

After choosing our relevant indicators, we created our analytic dataset using the difference between 2022 and 2009 values for each indicator. We used difference variables since we are estimating the rate of progress in percent minimum proficiency, where we are interested in the change over time, not just the starting minimum proficiency percentage. Changes in policy, social conditions, and other relevant conditions can lead to changes in the rate of growth in reading and mathematics proficiency (Kyriakides et al., 2017; Strietholt et al., 2019). At the time of this writing many countries did not have data for 2022, so we substituted 2021 data when necessary. A complete description of the indicators chosen, what years we used to form the difference variables, and the breakdown of Bayesian stacking models is contained in Table 2.

The first step of the workflow for this paper begins with estimating the rate of progress in the percentage of students meeting or exceeding minimum proficiency in reading and mathematics for both boys and girls separately. Our approach to the estimation of these growth curve models recognizes that minimum proficiency is a proportion that varies continuously across countries and therefore is not normally distributed as linear growth curve modeling assumes. Therefore, we address this problem by treating minimum proficiency as having been generated from a $beta(a, b)$ distribution, where a and b are shape and scale parameters, respectively (see e.g. Kaplan, 2023). The *beta* distribution is appropriate for data in the range [0,1].

In line with Kaplan et al. (2025), three models are estimated on the logit transformed data: (a) a linear growth curve model, (b) a latent basis model allowing the last latent basis term (2022) to be estimated by the data in order to examine the potential impact of the pandemic on the proportion of students in countries achieving minimum proficiency, and (c) a latent basis model that allows the last two latent basis terms (2018 and 2022) to be free in order to capture changes in the percentage achieving minimum proficiency that were already occurring prior to the pandemic. We evaluate each model using the LOOIC and KLD and choose one of these models for the next step in the workflow.

In the second step of the workflow, we adopt an $\mathcal{M}$-closed framework and utilize BMA to predict the percentage of students achieving or exceeding minimum proficiency. Again, because we need to account for the non-normality of minimum proficiency we apply BMA to the logit of the minimum proficiency values. For this analysis, we also examine the sensitivity of the results to changes in the initial starting conditions of the analysis by examining alternative model and parameter priors as discussed in Section 6.3 and Appendix A.

In the third step of the workflow, we will use the model averaged coefficients from the BMA prediction of the logit of rate-of-change to develop out-of-sample predictions of the percentage of students achieving minimum proficiency by plugging in the optimally predicted growth rate into our growth model and projecting the next two cycles of PISA, namely 2029 and 2033. At this point, we transform the predicted results back to 0-100 scale in order to obtain plots. Plots will be provided for the average across all countries and for each country separately. Four case study countries will be provided.

7.2 Missing Data Procedures

Given the numerous data sources, years of interest, and countries included in our analysis, we encountered some missing data due to some countries not collecting or reporting data for our chosen indicators. To maintain as large of a sample with as many relevant indicators as possible, our missing data imputation procedure is as follows.

Due to initial difficulty achieving convergence with our multiple imputation algorithm, when available, we substituted a neighboring year's observed data instead of solely relying on predictive mean matching for missing data. For example, the indicator "gross enrollment ratio primary school (both sexes)" only has reported data for eight of our selected 53 countries in 2022 at the time of this writing. Meanwhile, 50 of 53 countries reported data in 2021. The observed 2021 values were substituted for 2022 when available, and then we used multiple imputations to impute the rest. After the substitutions, only three countries required imputation for this indicator as opposed to 50 originally missing. This procedure ensured we could maximize the number of indicators included in our analysis. Despite participating in all 2009-2022 PISA cycles, Chinese Taipei was removed from our analytic dataset due to missing data on almost all the included indicators, which made multiple imputation algorithm convergence challenging. Once all available observed substitute data were used, we then imputed the remaining missing values using *predictive mean matching* (Rubin, 1986) via the R package mice (van Buuren & Groothuis-Oudshoorn, 2011).

Predictive mean matching produces unbiased estimates under the assumption that the data are missing completely at random (MCAR) or missing at random

(MAR) (Little & Rubin, 2019). In the present analysis, we assume the data are missing at random. Predictive mean matching imputes missing values by matching the predicted values from the observed data using a predictive mean metric to the predicted values with regression imputation. Next, the observed value is used for the imputation. So for each regression model, there is a predicted value for both observed and missing data. The predicted value for the observed data is then matched to the predicted missing data value, for example using a nearest neighbor metric. After a match is made, the missing value is replaced by the observed value (as opposed to the predicted value). Random matches are chosen when multiple matches are found.

To best account for uncertainty surrounding the imputation, multiple draws for missing values should be made. However, we only conducted this process for a single imputed data point to ensure model convergence. While using multiple imputations is considered best practice for producing the most reliable results, we argue that using a single imputation via predictive mean matching is a significant improvement of listwise deletion or using a much smaller set of complete indicators.

8 Results

This section presents the results in alignment with the workflow described in Section 7.

8.1 Growth modeling results

Table 3 presents growth curve modeling results under three specifications: (a) the linear model (M0), (b) latent basis model (M1) wherein the last time point is freely estimated, allowing a data-based estimation of the potential impact of COVID, and (c) latent basis model 2 (M2) which allows the last two time points to be estimated under the assumption that the decline in the percentage reaching minimum levels of proficiency occurred before the onset of the pandemic.

Table 3: Posterior estimate of starting percent minimum proficiency and rates of progress, 95% credible intervals (in parentheses), and predictive evaluation under linear and two latent basis models.

	Reading		Mathematics	
	Boys	*Girls*	*Boys*	*Girls*
Starting minimum proficiency percentage				
Linear model M0	68.24 (65.05, 71.51)	82.45 (80.08, 84.67)	70.97 (67.57, 74.07)	69.93 (66.27, 73.16)
Latent basis M1	68.22 (64.97, 71.41)	82.44 (79.90, 84.54)	70.99 (67.70, 74.30)	69.86 (66.36, 73.25)
Latent basis M2	68.23 (65.05, 71.38)	82.44 (80.00, 84.54)	71.09 (67.80, 74.25)	69.98 (66.60, 73.48)
Rate of progress				
Linear model M0	-1.10 (-2.11, -0.11)	-4.07 (-5.11, -3.02)	-1.30 (-2.37, -0.19)	-0.17 (-1.63, 1.30)
Latent basis M1	-1.10 (-2.08, -0.06)	-4.00 (-4.97, -3.01)	-1.40 (-2.51, -0.31)	0.16 (-1.69, 1.84)
Latent basis M2	-1.09 (-2.05, -0.05)	-4.02 (-5.14, -2.96)	-1.87 (-3.03, -0.62)	-0.31 (-2.45, 1.71)
Predictive evaluation				
Linear model M0	215.57	259.52	207.30	237.58
Latent basis M1	215.68	258.55	207.08	237.43
Latent basis M2	215.77	259.44	203.99	239.32

Predictive performance is compared using LOO-IC estimates for each condition.

For boys and girls both we observed virtually no difference between the models according to the LOO-IC predictive criterion. An inspection of the estimated starting percent minimum proficiency and rate of progress also showed no substantive differences, with the possible exception of the rate of progress in mathematics minimum proficiency for girls under latent basis M1 and M2 relative to the linear model. For the remainder of the paper, we will use estimates of the rate of progress in meeting minimum proficiency under latent basis M1, again to allow for any possible impact of the pandemic.

8.2 Bayesian model averaging results

Bayesian model averaging was applied to the 31 predictors shown in Table 2 using the R program BMS (Zeugner & Feldkircher, 2015). The analysis utilized unit information

priors for all model parameters and uniform model priors. To begin, Table 4 presents the posterior model probabilities (PMPs) for the top 5 models explored by the BMA algorithm as well as the total PMP across all models explored in the model space.

Table 4: Posterior model probabilities for the top five models along with total posterior model probability across all models.

	Reading		**Mathematics**	
	Boys	*Girls*	*Boys*	*Girls*
Model 1	0.04	0.03	0.03	0.04
Model 2	0.02	0.02	0.02	0.02
Model 3	0.01	0.02	0.02	0.02
Model 4	0.01	0.02	0.02	0.02
Model 5	0.01	0.02	0.02	0.02
Total PMP	*0.54*	*0.47*	*0.56*	*0.54*

We find that the total PMPs for both reading and mathematics and across boys and girls are well below 1.0 suggesting substantial model uncertainty and thus justifying the use of BMA over the selection of one specific model. Indeed the top model (Model 1) which would be the model selected on the basis of the Bayesian information criteria illustrates the extent of model uncertainty.

Tables 5 and 6 present the results for the top 6 predictors of the rate of progress in the percentage of boys and girls at or above the PISA minimum proficiency for reading and mathematics, respectively. The column labeled PIP shows the posterior inclusion probabilities, which represents the proportion of the total posterior model probability (or model mass) that includes that predictor across all models considered in the averaging process. For example, we find that for boys, the top predictor of the rate of progress in reading minimum proficiency is the change from 2009 to 2022 (or 2021) in the percentage of teachers in pre-primary education who are female. However, the PIP is 0.78, meaning that 78% of the total posterior model probability mass includes that predictor. It is arguble whether this predictor is important in a substantive sense. Contrast this with the rate of progress in minimum proficiency for girls in mathematics. Here we find that 96% of the total probability mass contains the predictor measuring the change from 2009 to 2022 (2021) in the percent of children out of school of primary school age. Arguably, this predictor is important for assessing the rate of progress in mathematics minimum proficiency for girls. This latter finding notwithstanding, across domains and for boys and girls we find that few of the predictors used in this study are clearly important in predicting the logit of the rate of progress in reading and mathematics. Although we believe that we have captured a considerable number of the most important predictors of the rate of progress in minimum proficiency driven by our understanding of the CIPO model, there is considerable room for future theory and indicator development that might improve PIP levels.

9 Overall country projections and case studies

We begin by presenting BMA-based projections for reading and mathematics and for boys and girls separately across all countries/economies used in this analysis.

9.1 Overall country projections

Overall country projections are shown in Figure 4.

We find that across the countries used in this analysis, there has been relatively small amounts of change in the percentage of students who have reached or exceeded the PISA-defined minimum proficiency level and that, moreover, the projection to 2033 appears to be relatively stable at about 65% at or above minimum proficiency across these countries. Another way to consider this finding is that without

Table 5: BMA results for reading proficiency. Note that the outcome of interest is the estimated rate of progress (i.e. growth rate) in the percentage of students at or above the PISA minimum proficiency. All indicators are measured as the difference between 2022 (or 2021) and 2009 values.

Boys	
	PIP
Percentage of teachers in pre-primary education who are female	0.78
Rural population percent of total population	0.52
Percentage of full-time teachers per school	0.43
Teaching hours upper-secondary	0.36
Teaching hours primary	0.31
Employment females as percent of employment annual	0.30

Girls	
	PIP
Employment females as percentage of employment annual	0.56
Gender development index	0.51
Percentage of teachers in pre-primary education who are female	0.50
Children out of school of primary age	0.41
Teaching hours upper-secondary	0.40
Official entrance age to primary education years	0.38

PIP: Posterior inclusion probability.

Table 6: BMA results for mathematics proficiency. Note that the outcome of interest is the estimated rate of progress (i.e. growth rate) in the number of students at or above the PISA minimum proficiency. All indicators are measured as the difference between 2022 (or 2021) and 2009 values.

Boys	
	PIP
Children out of school of primary-school age	0.90
Index highest occupational status of parents	0.67
Percentage of full-time teachers per school	0.66
Employment females as percentage of employment annual	0.46
Percentage of pre-primary teachers who are female	0.45
Teaching hours lower secondary	0.41

Girls	
	PIP
Children out of school of primary-school-age	0.96
Employment females as percentage of employment annual	0.48
Teaching hours upper-secondary	0.42
Index school size	0.42
Percentage of part-time teachers per school	0.40
Index highest occupational status of parents	0.38

PIP: Posterior inclusion probability

intervention, an average of approximately 35% of 15-year-old students across these countries will remain below minimum proficiency in reading and mathematics beyond 2030, independent of gender.

9.2 Case studies

In this section, we present the results of four case studies: Croatia, Germany, the United Kingdom, and the United States. These countries were chosen to illustrate the differences in the predictive quality of the model across boys and girls and across

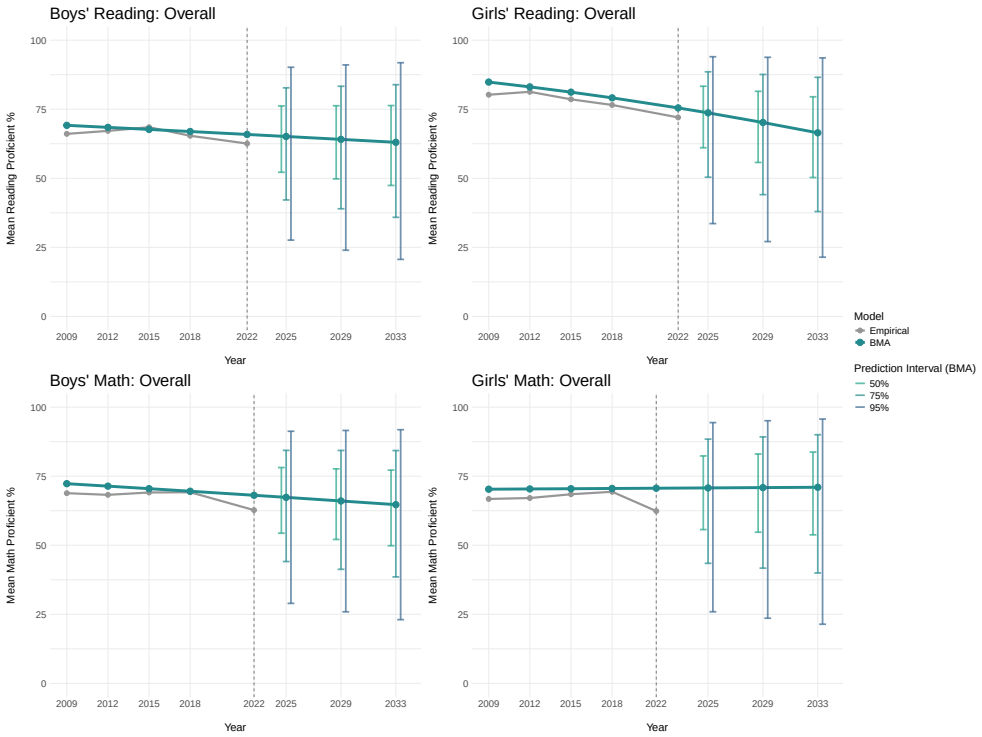

Fig. 4: BMA-based projections and prediction intervals for mathematics (a) and reading (b) for boys and girls. Note that now PISA is on an every four-year cycle.

reading and mathematics. Remaining countries are provided in the supplementary material.

First, to assess the quality of the prediction models for our case studies, we utilize *BMA predictive densities* as implemented in BMS (Zeugner & Feldkircher, 2015). These predictive densities provide a means of checking the prediction model of the growth rate based on BMA against the unconditional growth rate based on latent basis M1, allowing examination of problems with the prediction model and/or potential outlier countries.[8] In the upper panels of the figures, the dashed vertical line is the unconditional growth rate and the solid line is the model-predicted (expected) growth rate based on the predictor information from the country of interest and the information from all other countries. Thus, our approach borrows information from all the countries in providing a prediction of the growth rate for the country of interest. The 95% quantiles for the predictive densities are also displayed.

Second, in the lower panel of the following figures, we present the trajectory plots of minimum proficiency for the case study countries, and for boys and girls and proficiency domains. The gray line is the empirical percentage of the countries students at or above the PISA level 2 minimum proficiency and the green line is the estimated trajectory under latent basis model M2, with the estimated intercept and growth rate obtained from BMA. We also describe change in the percentage of students at or above minimum proficiency from 2009 to the projected value in 2033 and these are show in Table 7 below.

9.3 Case study 1: Croatia

Predictive density plots and trend plots for Croatia are displayed in Figure 5. For boys and girls and for both proficiency domains, the BMA predictive growth rates are all within the 95% quantiles of the unconditional growth rate, however, we find that the fit is not excellent, particularly for boys in reading and for girls in reading and mathematics.

[8]This analysis assumes that latent basis model M1 is very close to the actual growth rate. This might be a reasonable assumption given the data based fitting of M1 and the fact that this model did not differ compared to the linear model and latent basis model M2.

Table 7: Projections of the percent of students meeting minimum proficiency for the case study countries by gender and literacy domain.

	Reading					
	Boys			*Girls*		
	2009	2033	% Change	2009	2033	% Change
Croatia	68	59	-9	88	66	-22
Germany	76	70	-6	87	67	-20
United Kingdom	77	76	-1	86	80	-6
United States	79	66	-13	86	68	-18
	Mathematics					
	Boys			*Girls*		
	2009	2033	% Change	2009	2033	% Change
Croatia	69	62	-7	65	64	-1
Germany	83	71	-12	80	73	-7
United Kingdom	82	75	-7	77	82	+5
United States	80	59	-21	73	67	-6

Note that 2033 is a projection based on the latent basis model M1.

The trajectory plots reveal that the trend is mostly constant for all cases except for a slightly more noticeable decline for girls' in the minimum proficiency percentage in reading over time. Inspecting Table 7, in 2009, the minimum proficiency percentage for Croatian boys in reading was 68% and in 2033 it is projected to be 58%, a drop of approximately 9%. For Croatian girls in reading, the projected decline is much greater with 88% of the girls at or above minimum proficiency in 2009 but projected to be at 66% by 2033 for a drop of 22%. Thus while the percent of Croatian girls at minimum proficiency is uniformly greater than that for boys, the decline is much greater for girls.

The minimum proficiency percentage for Croatian boys in mathematics is projected to drop by 7% in 2033 compared to where it was in 2009. For Croatian girls in mathematics, the projected decrease in the minimum proficiency percentage is much smaller at 1%.

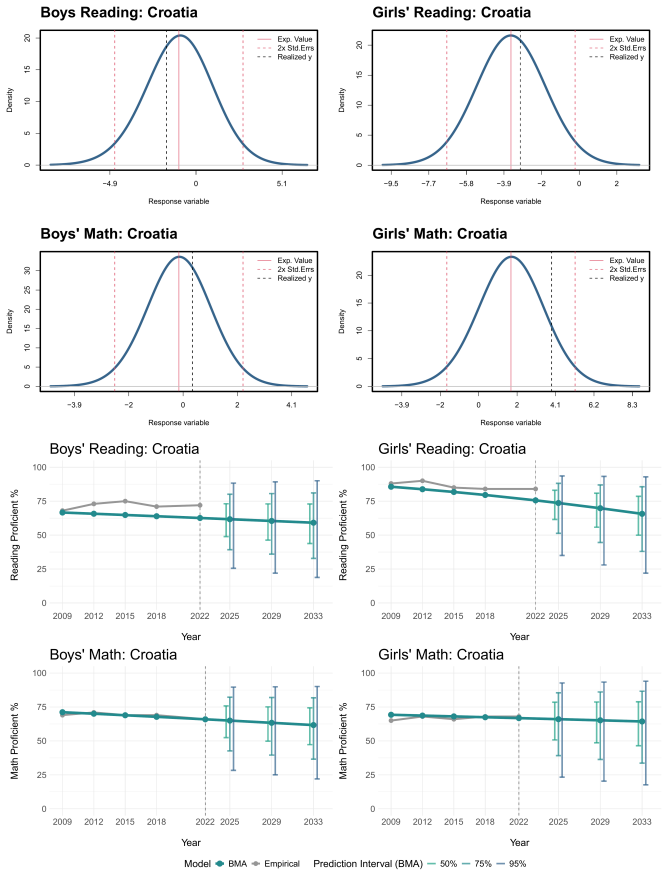

Fig. 5: Predictive density plots (above) and trend lines (below) for percentage of Croatian students at the minimum proficiency level in mathematics from 2009 to 2022.

9.4 Case study 2: Germany

Predictive density plots and trend plots for Germany are displayed in Figure 6. For math, model fit was quite good for boys and girls, whereas for reading the model fit was not as good. Nevertheless, in all four cases, the BMA predicted rate of progress was within the 95% prediction interval of the actual growth rate.

Inspection of the trajectory plots for Germany as well as Table 7 reveal a more noticeable decline in the percent of German girls meeting or exceeding minimum proficiency in reading compared to German boys. In 2009, the percent of German boys reaching or exceeding minimum proficiency in reading was 76% and projected to be 70% in 2033 for a drop of 6%. In contrast, although the percent of German girls reaching or exceeding minimum proficiency is greater than German boys in 2009, the decline is much greater at 20%.

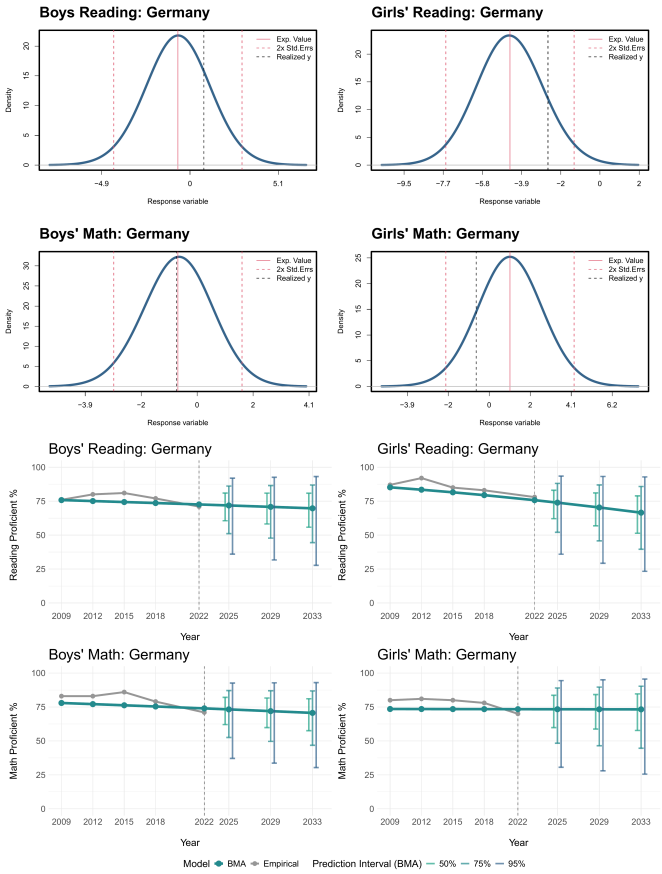

Fig. 6: Predictive density plots (above) and trend lines (below) for percentage of German students at the minimum proficiency level in mathematics from 2009 to 2022.

9.5 Case study 3: United Kingdom

Predictive density plots and trend plots for the United Kingdom are displayed in Figure 8. From Table 7, the percentage of UK boys reaching or exceeding minimum proficiency for reading in 2033 is predicted to be 76%, a 1% drop compared to 2009. For UK girls, the projected percentage meeting or exceeding minimum proficiency in 2033 is 80% but representing a larger drop of 6% from 2009 compared to boys.

Regarding mathematics, the percent of UK boys meeting or exceeding minimum proficiency in 2033 is 75% representing a 7% drop from the 2009 baseline percentage. For UK girls, we predict a 5% increase.

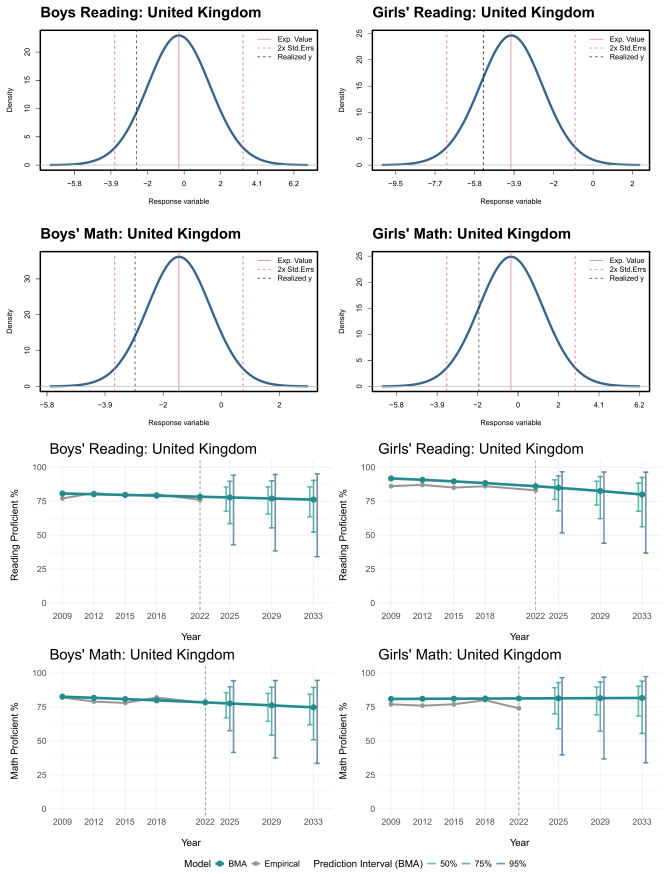

Fig. 7: Predictive density plots (above) and trend lines (below) for percentage of UK students at the minimum proficiency level in mathematics from 2009 to 2022.

9.6 Case Study 4: United States

Finally, in Table 7 for the United States, the projected percentage of boys meeting or exceeding minimum proficiency in reading in 2033 is 66% representing a 13% drop compared to the 2009 baseline of 79%. For US girls, the projected percentage meeting or exceeding minimum proficiency is 68% representing an 18% decline relative to the baseline 2009 percentage. This decline is despite the fact that girls start off at an advantage over boys in reading and maintain that advantage in 2033.

Regarding mathematics, we find that the percent of US boys projected to meet or exceed minimum proficiency in 2033 is 59% representing an 21% decline from the 2009 baseline. We find that the percent of US girls projected to meet or exceed minimum proficiency in 2033 67% representing a lesser decline of 6% from the 2009 baseline.

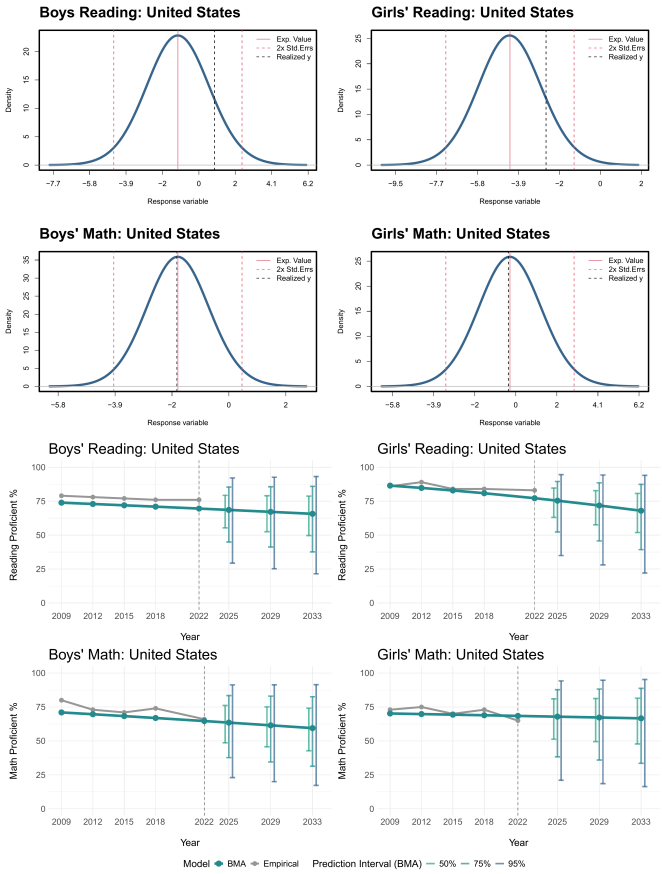

Fig. 8: Predictive density plots (above) and trend lines (below) for percentage of US students at the minimum proficiency level in mathematics from 2009 to 2022.

10 Conclusions

Using country-level trend data from PISA, this paper provided probabilistic projections of the percent of students in a country meeting PISA-defined minimum proficiency levels in reading and mathematics. Our probabilistic projections relied on a novel combination of Bayesian growth curve modeling as well as methods for addressing model uncertainty via BMA, and relying on a large set of predictors derived from OECD, UNESCO, and World Bank data sources. Our approach was based on obtaining an optimal prediction of the PISA 2009 starting minimum proficiency percentage and the rate of change in these percentages until 2022 and then plugging these estimates into a growth model to predict minimum proficiency percentage to 2033.

A benefit of BMA is that it provides a measure of variable importance on the basis of posterior inclusion probabilities. However, our results revealed that although one or two variables were deemed important across the reading and mathematics domains for boys and girls, their impact on the rate of progress in minimum proficiency was quite modest. This finding speaks to the need to develop parallel theory as to how changes in system level variables might impact change over time in minimum proficiency. Developing such theory was beyond the scope of this paper.

A clear limitation of this work, and thus an opportunity for future research, concerns the sample of countries used in this analysis. Specifically, for this study, we relied on data from OECD countries or those that partnered with the OECD to conduct PISA, as well as selecting those countries that provided data from 2009 to 2022. Missing from this study are data from countries that did not participate in PISA. An important line of research going forward would be to harmonize data from ILSAs such as PISA with existing regional assessments such as the ERCE (UNESCO

& LLECE, 2020; UNESCO-OREALC, 2023), PASEC (CONFEMEN - Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN, 2020), PILNA (Pacific Community, 2021), SACMEQ (Southern & for Monitoring Educational Quality (SACMEQ), 2019), and SEA-PLM (UNICEF et al., 2019), to provide additional information not fully covered by the ILSAs, and indeed, countries making up these regional assessments are often (but not in all cases) those where noticeable inequities in literacy and numeracy are located. This would be no small task insofar as these assessments have different goals, sampling designs, and assessment instruments. Regardless, even if harmonization across these international and regional assessments were successful, there are still many countries in the world with either very poor or non-existent data for monitoring minimum proficiency in reading and mathematics, making it extremely difficult to provide projections to 2030 (see e.g. Nilashi et al., 2023).

With the above caveats and limitations in mind, we conclude that for the countries examined in this study, and with the novel approaches of combining growth models while addressing model uncertainty, all within a Bayesian perspective, that without interventions tailored to the needs of the individual countries, it seems unlikely that countries will reach their SDG 4.1.1c targets by 2030 and that the percentage of students at or above minimum proficiency in reading and mathematics will continue to remain stable or slowly decline as it has in the past. As a necessary driver for reaching many of the SDGs, quality education and providing equal learning opportunities for all must be an urgent focus of education policy efforts around the globe.

References

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

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.

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

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

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models : A structural equation perspective*. New York: John Wiley.

Borgonovi, F., Ferrara, A., & Maghnouj, S. (2018). *The gender gap in educational outcomes in norway* (Tech. Rep.). Paris: OECD. (OECD Education Working Paper No. 183)

Borgonovi, F., Pokropek, A., Jakubowski, M., & Piacentini, M. (2017). *Youth in transition: How do some of the cohorts participating in pisa fare in piaac?* (OECD Education Working Papers No. 155). Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/51479ec2-en doi: 10.1787/51479ec2-en

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

Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*(5), 1986–2018. doi: 10.1214/15-AOS1334

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

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

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

CONFEMEN - Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN. (2020). *PASEC 2019 International Report: Quality of Education Systems in French-speaking Sub-Saharan Africa: Teaching/Learning Performance and Environment in Primary Education* (Tech. Rep.). PASEC, CONFEMEN. (https://pasec.confemen.org/en/ressource/diagnostic-evaluation/)

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

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

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

Encinas-Martín, M., & Cherian, M. (2023). *Gender, education and skills: The persistence of gender gaps in education and skills*. Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/34680dd5-en doi: 10.1787/34680dd5-en

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

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

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

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

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*, 107–114.

Grimm, K. J., Ram, N., & Estabrook, R. (2017). *Growth modeling: Structural equation and multilevel modeling approaches*. New York: Guilford.

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

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. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika*, *86*, 215–238.
Kaplan, D. (2023). *Bayesian statistics for the social sciences* (2nd ed.). New York: Guilford Press.
Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*, 505–517.
Kaplan, D., Harra, K., Stampka, J., & Jude, N. (2025). Stacking models of growth: A methodology for predicting the pace of progress to the education sustainable development targets using international large-scale assessments. *Psychometrika*, 1-29.
Kaplan, D., & Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: a case study using naep. *Large-scale Assessments in Education*, *9*, 1-15. Retrieved from https://doi.org/10.1186/s40536-021-00108-2 doi: 10.1186/s40536-021-00108-2
Kaplan, D., & Lee, C. (2015). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling*, 1-11.
Kaplan, D., & Lee, C. (2018). Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review*. doi: 10 .1177/0193841X18761421
Kaplan, D., & Yavuz, S. (2019). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research*. Retrieved from https://doi.org/10.1080/00273171.2019.1657790 (PMID: 31538505) doi: 10.1080/00273171.2019.1657790
Kass, R. E., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343–1370.
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.
Kyriakides, L., Georgiou, M. P., Creemers, B. P. M., Panayiotou, A., & Reynolds, D. (2017). The impact of national educational policies on student achievement: a european study. *School Effectiveness and School Improvement*, *29*(2), 171–203. Retrieved from https://doi.org/10.1080/09243453.2017.1398761 doi: 10.1080/09243453.2017.1398761
Lafortune, G., Fuller, G., Moreno, J., Schmidt-Traub, G., & Kroll, C. (2018). *SDG index and dashboards: Detailed methodological paper* (Tech. Rep.). Sustainable Development Solutions Network (SDSN). (https://raw.githubusercontent.com/sdsna/2018GlobalIndex/master/\2018GlobalIndexMethodology.pdf)
Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data* (Vol. 53). New York: Wiley.
Levin, B. (2003). *Approaches to equity in policy for lifelong learning* (Tech. Rep.). OECD. https://www.oecd.org/education/school/38692676.pdf. (A paper commissioned by the Education and Training Policy Division, OECD, for the Equity in Education Thematic Review)
Ley, E., & Steel, M. F. J. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, *24*, 651–674.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.
Lindley, D. V. (1991). *Making decisions*. London: Wiley.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd. ed.). John Wiley & Sons.

Madigan, D., & Raftery, A. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*, 1535–1546.

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

Montoya, S., & MacDonald, K. (2022). *Monitoring of the sustainable development goals using large-scale international assessments: A strategy for reporting sdg 4 indicators using data from cross-national assessments* (Tech. Rep.). UNESCO Institute for Statistics. Retrieved from https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/04/Monitoring-of-the-SDGs-Using-Large-Scale-International-Assessments_April-2022.pdf

Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology*, *27*, 453–480.

NCES. (2022). *National assessment of educational progress (NAEP).* https://nces.ed.gov/nationsreportcard/.

Nilashi, M., Keng Boon, O., Tan, G., Lin, B., & Abumalloh, R. (2023). Critical data challenges in measuring the performance of sustainable development goals: Solutions and the role of big-data analytics. *Harvard Data Science Review*, *5*. Retrieved from https://hdsr.mitpress.mit.edu/pub/9n4uzkg3 (Accessed June 24, 2025)

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris: Organization for Economic Cooperation and Development.

OECD. (2008). *Ten steps to equity in education. policy brief.* Paris: Author.

OECD. (2018). *Equity in education: Breaking down barriers to social mobility* (Tech. Rep.). Paris: Author. https://doi.org/10.1787/9789264073234-en.

Pacific Community. (2021). *Pacific Islands Literacy and Numeracy Assessment (PILNA) 2021 Technical Report* (Tech. Rep.). Suva, Fiji: Educational Quality and Assessment Programme (EQAP). (Available at https://pacificdata.org/data/dataset/pilna-2021-regional-report)

Porwal, A., & Raftery, A. E. (2022). Comparing methods for statistical inference with model uncertainty. *Proceedings of the National Academy of Sciences*, *119*(16), e2120737119. Retrieved from https://www.pnas.org/doi/10.1073/pnas.2120737119 doi: 10.1073/pnas.2120737119

Porwal, A., & Raftery, A. E. (2023). Effect of model space priors on statistical inference with model uncertainty. *New England Journal of Statistics in Data Science*, *4*(1), 1–20. Retrieved from https://nejsds.nestat.org/journal/NEJSDS/article/16/read doi: 10.51387/22-NEJSDS16

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raftery, A., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics*, *4*, 87–95.

Scheerens, J. (2023). Theory on teaching effectiveness at meta, general and partial level. In A.-K. Praetorius & C. Y. Charalambous (Eds.), *Theorizing teaching.* Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-031-25613-4_4 doi: 10.1007/978-3-031-25613-4_4

Schleicher, A. (2008). Student learning outcomes in mathematics from a gender perspective: What does the international pisa assessment tell us? In M. Tembon & L. Fort (Eds.), *Girls? education in the 21st century. gender equality, empowerment, and economic growth* (pp. 41–52). Washington: The World Bank.

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

Sloughter, J. M., Gneiting, T., & Raftery, A. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*,

141, 2107–2119.
Southern, & for Monitoring Educational Quality (SACMEQ), E. A. C. (2019). *Sacmeq iv: The quality of education in southern and eastern africa* (Tech. Rep.). SACMEQ.
Statisticat, & LLC. (2021). Laplacesdemon: Complete environment for bayesian inference [Computer software manual]. Bayesian-Inference.com. Retrieved from https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software (R package version 16.1.6)
Strietholt, R., et al. (2019). The impact of education policies on socioeconomic inequality in student achievement: A review of comparative studies. In L. Volante, S. Schnepf, J. Jerrim, & D. Klinger (Eds.), *Socioeconomic inequality and student outcomes* (Vol. 4). Singapore: Springer. Retrieved from https://doi.org/10.1007/978-981-13-9863-6_2 doi: 10.1007/978-981-13-9863-6_2
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.
UN General Assembly. (2015). *The 2030 agenda for sustainable development.* https://www.refworld.org/docid/57b6e3e44.html. (Accessed 15 June 2019.)
UNDP. (2020). *From insights to action: Gender equality in the wake of COVID-19.* New York, NY: United Nations Development Program. Retrieved from https://www.unwomen.org/en/digital-library/publications/2020/09/gender-equality-in-the-wake-of-covid-19
UNESCO. (2015). *Education for all global monitoring report 2015: Achievements and challenges.* Paris: Author.
UNESCO. (2016a). *Education 2030: Incheon declaration and framework for action towards inclusive and equitable quality education and lifelong learning for all.* https://unesdoc.unesco.org/ark:/48223/pf0000245656. (Accessed: 2024-08-13)
UNESCO. (2016b). *Global education monitoring report. gender review: Creating sustainable futures for all.* Paris: Author.
UNESCO. (2018). *Handbook on measuring equity in education.* Montreal: UNESCO Institute for Statistics.
UNESCO, & LLECE. (2020). *Análisis curricular del erce 2019 del conjunto de países que conforman la cecc/sica* (Tech. Rep.). UNESCO. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000375368
UNESCO-OREALC. (2023). *Reporte técnico. cuarto estudio regional comparativo y explicativo (erce 2019)* (Tech. Rep.). Author.
UNICEF, of Education Organization (SEAMEO), S. A. M., & for Educational Research (ACER), A. C. (2019). *Sea-plm 2019 main regional report: Quality of education in southeast asia* (Tech. Rep.). UNICEF, SEAMEO, ACER. (https://research.acer.edu.au/ar_misc/52/)
United Nations. (2015). *The millennium development goals report 2015.* New York, NY: United Nations Educational, Scientific and Cultural Organization (UNESCO). https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf.
van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1-67.
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. doi: 10.1007/s11222-016-9696-4
Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. Retrieved from DOI:10.1214/12-SS102
von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., ... Yin, L. (2024). *Timss 2023 international results in mathematics and science.* Retrieved from https://doi.org/10.6017/lse.tpisc.timss.rs6460 doi: 10.6017/lse.tpisc.timss.rs6460
Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely

applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363–381.

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

World Bank, UNICEF, UNESCO, FCDO, USAID, & Bill & Melinda Gates Foundation. (2022). *The state of global learning poverty: 2022 update.* Retrieved from https://www.unicef.org/media/122921/file/State%20of%20Learning%20Poverty%202022.pdf (Conference Edition, June 23, 2022)

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

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti. Studies in Bayesian econometrics* (pp. 233–243). New York: Elsevier.

Zeugner, S., & Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, *68*(4), 1–37. doi: 10.18637/jss.v068.i04