A Bayesian Model for Estimating Sustainable Development Goal Indicator 4.1.2: School Completion Rates

Dharamshi, Ameer; Barakat, Bilal; Alkema, Leontine; Antoninis, Manos

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Dharamshi, A., Barakat, B., Alkema, L., & Antoninis, M. (2022). A Bayesian Model for Estimating Sustainable Development Goal Indicator 4.1.2: School Completion Rates. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *71*(5), 1822-1864. https://doi.org/10.1111/rssc.12595

Received: 21 September 2021 | Accepted: 11 August 2022

DOI: 10.1111/rssc.12595

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series C Applied Statistics

A Bayesian model for estimating Sustainable Development Goal indicator 4.1.2: School completion rates

Ameer Dharamshi[1] | **Bilal Barakat**[2] | **Leontine Alkema**[3] | **Manos Antoninis**[2]

[1]University of Toronto, Toronto, Ontario, Canada

[2]Global Education Monitoring Report, UNESCO, Paris, France

[3]University of Massachusetts Amherst, Amherst, Massachusetts, USA

Correspondence
Ameer Dharamshi, University of Toronto, Toronto, ON, Canada.
Email: ameer.dharamshi@mail.utoronto.ca

Abstract

Estimating school completion is crucial for monitoring Sustainable Development Goal (SDG) 4 on education. The recently introduced SDG indicator 4.1.2, defined as the percentage of children aged 3–5 years above the expected completion age of a given level of education that have completed the respective level, differs from enrolment indicators in that it relies primarily on household surveys. This introduces a number of challenges including gaps between survey waves, conflicting estimates, age misreporting and delayed completion. We introduce the Adjusted Bayesian Completion Rates (ABCR) model to address these challenges and produce the first complete and consistent time series for SDG indicator 4.1.2, by school level and sex, for 164 countries. Validation exercises indicate that the model appears well-calibrated and offers a meaningful improvement over simpler approaches in predictive performance. The ABCR model is now used by the United Nations to monitor completion rates for all countries with available survey data.

KEYWORDS

Bayesian modelling, household surveys, misreporting, school completion, SDG 4

1 | INTRODUCTION

The 2030 Agenda for Sustainable Development, adopted in 2015, established a framework for global progress centred around the 17 Sustainable Development Goals (SDG). The SDGs are a broad set of objectives focused on 'people, planet and prosperity'. They include SDG 4 on education, defined to 'ensure inclusive and equitable quality education and promote lifelong learning opportunities for all' (United Nations, 2015).

Historically, measuring progress in education development has focused on enrolment and attendance statistics. Recently, however, SDG 4 has shifted attention away from mere enrolment towards completion and learning, notably through target 4.1, which calls on countries to 'ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes'. In 2015, the proposal for SDG global indicator 4.1.1, defined as the 'proportion of children and young people (a) in Grade 2 or 3; (b) at the end of primary education; and (c) at the end of lower secondary education achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex', reflected this shift (UN Statistical Division, 2021a). But the monitoring framework was deficient as it focused on the learning outcomes of the student population without providing information on the percentage of children and youth population that reach the end of each education cycle. A true measure of progress requires the monitoring of the education trajectories of all children and youth.

Just as it had called in 2012 on a learning outcome indicator for the post-2015 agenda (UNESCO, 2012), the 2016 Global Education Monitoring (GEM) Report called for a measure of completion to be added to the monitoring of SDG target 4.1 (UNESCO, 2016). Such a measure would be preferably based on survey data to overcome challenges associated with administrative data that had prevented a reliable assessment of education progress from emerging. The UNESCO Institute for Statistics, as custodian agency of almost all SDG 4 indicators, made the case, and the Inter-agency and Expert Group on SDG Indicators adopted the completion rate at three levels of education (primary, lower secondary and upper secondary) as SDG global indicator 4.1.2, one of only six successful out of more than 200 proposals made during the 2020 Review. The new indicator's metadata foresaw a methodological development to address its limitations (UN Statistical Division, 2021b). This paper describes the methodology that has since been developed and is being used to report on the indicator (United Nations, 2022).

Indicator 4.1.2 is defined as the 'percentage of a cohort of children or young people aged 3-5 years above the intended age for the last grade of each level of education who have completed that grade' (UN Statistical Division, 2021b). In other words, the completion rate is a 'flow' measure of attainment, aiming to capture the peak of education development progress in the current generation, unlike 'stock' measures of attainment that have tended historically to focus on adult cohorts (Barro & Lee, 1993). While a stock measure is relevant for analyses of countries' growth, a flow measure allows for a timely reading of progress, facilitating education policy responses.

Traditionally, education monitoring has relied on administrative enrolment records. But the 2030 Agenda for Sustainable Development has helped shift emphasis to the use of survey data. This is mainly due to the focus on disaggregation to capture the spirit of 'leaving no one behind'. But administrative data are also often not suitable to measure completion. For instance, they do not allow easy distinctions to be made between repeaters and non-repeaters, especially in poorer countries. Data on graduates are often not available, especially at the upper secondary education level, where education pathways are more fragmented, while this information tends not to be available by age. These concerns are on top of the usual concern over the accuracy of population measures.

Poor understanding of the weaknesses of the administrative data-equivalent indicator, the gross intake rate to the last grade of primary school, has led in the past to wrong conclusions. For instance, the World Bank claimed in 2011 that countries such as Myanmar and the United Republic of Tanzania had achieved universal primary completion (World Bank, International Monetary Fund, 2011), when in fact at least one in four and one in five children, respectively, were not reaching the end of primary school.

This shift towards mainstreaming survey and census data in education monitoring is in line with the calls for a data revolution post-2015, which stressed that 'the more data can be combined, the more useful they are' (Independent Expert Advisory Group on the Data Revolution for Sustainable Development, 2014).

Survey data, however, bring their own challenges. Most cross-country comparable household survey programmes conduct a survey in a given country at most every 3 to 5 years and the results released at least 1 year later, generating a considerable time lag. For most countries, multiple surveys are available though they may provide conflicting information. The 2016 GEM Report raised the question of reconciling the different sources (UNESCO, 2016). Simply averaging estimates or fitting a standard linear regression trend ignores relevant information. Some sources may show greater variability due to small sample size or other, non-statistical issues that make them less reliable. By itself, this could be accounted for using weighted linear regression. This method still does not recognise, however, that some sources may systematically result in lower or higher estimates relative to others. Such bias can reflect differences in sampling frames or how questions are asked. In addition, some respondents provide information retrospectively and the time that has lapsed increases the risk of errors that need to be corrected.

Figure 1 illustrates the problem using primary completion rates. It is clear that an assessment of the trend has to consider the relatively larger uncertainty of the 2003 estimate, and that recent Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) surveys systematically differ in their baseline. Comparing them directly, or always adopting the 'latest available' as the best estimate would lead to the conclusion that there have been large jumps in completion in a short amount of time, in opposite directions. These challenges also constrain the calculation of consistent trends in completion rates for regional aggregates because the set of countries with observations in a given year changes over time.

The international health community faced similar challenges in measuring indicators based on multiple sources, such as under-5 mortality or maternal mortality rates (MMR). The UN Inter-agency Group for Child Mortality Estimation adopted a consensus model to generate annual estimates for under-5 (Alkema & New, 2014) and neo-natal mortality (Alexander & Alkema, 2018) in each member state. The Inter-Agency Group for Maternal Mortality Rates followed a similar process (Alkema et al., 2016).

In this paper, we introduce the Adjusted Bayesian Completion Rates (ABCR) model, which takes inspiration from the general approach to estimating health indicators using Bayesian hierarchical models, but fully adapts them to the education context. Structurally, the ABCR model has similarities with the above mortality models in that it estimates an underlying trend in target values and shares information on parameter scaling across countries (Alexander, 2020). However, given the substantial differences in data considerations and broader context, the ABCR model proposes a new process to maximise survey data utilisation, as well as education-specific correction terms.

Specifically, we address concerns about limited data by introducing a new process to reconstruct historical completion rates from the available surveys and recognise the increased error associated with retrospective series. We also explicitly model late completion by specifying the

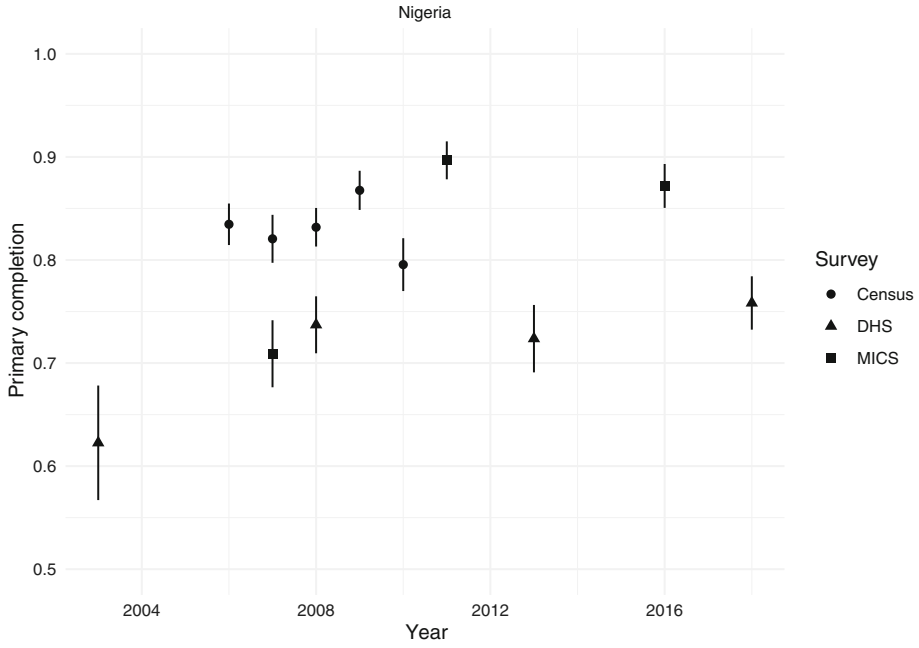

FIGURE 1 Primary school completion rate in Nigeria 5 years above nominal age for final primary grade, from different surveys, with estimated 95% uncertainty intervals. N.B. *y*-axis starts at 0.5.

delay magnitude as a function of age, and address age-misreporting concerns stemming from limited survey respondent numeracy skills in some regions. Such adjustments permit the ABCR model to consolidate survey data into a smooth underlying trend in completion rates from which the estimated true annual completion rates for each country can be extracted. The ABCR model is now used by the United Nations to estimate SDG indicator 4.1.2 and monitor regional and global progress.

The ABCR model is the first to estimate completion rates as defined by SDG global indicator 4.1.2, as opposed to adult educational attainment rates commonly used for other purposes. The model estimates the official completion rate, which focuses on the age group 3 to 5 years above the official graduation age to capture the late completion phenomenon, which is very common in low- and lower-middle-income countries. In addition to the official completion rate indicator definition, the model also estimates an 'ultimate' completion rate that captures very delayed education trajectories that characterise some countries, usually the detriment of girls' education progress. By addressing the various data quality concerns associated with survey data, these estimates are also less sensitive to individual surveys, the year in which they were conducted, and the type of survey that happens to be the latest available in a given country.

We begin this paper by formally establishing the official definition of the completion rate indicator and the associated notation before presenting the structure of the model and how it addresses the specific challenges posed by estimating this indicator. We then present a first set of estimates along with an assessment of their quality and performance against more simplistic approaches. Finally, we conclude with a summary of the outcomes and challenges associated with the ABCR model. The analyses in this paper are based on a consolidated collection of 696 microlevel datasets on school completion from 164 countries. The sources are described in detail in Appendix B.1.

2 | DEFINITIONS AND DATA

2.1 | Completion rates

SDG global indicator 4.1.2 measures completion among individuals who are between 3 and 5 years above the theoretical age for the final grade of the education level in question (UN Statistical Division, 2021b). This theoretical final grade age is the age of a child who starts school at the official school entry age and progresses one grade each year. We refer to this age as a_0. Then, we refer to age $a_0 + 3$ as a_3, $a_0 + 5$ as a_5, and generally to $a_0 + n$ as a_n. Note that the age value corresponding to a_0 is defined by a given school system and thus varies across countries. In a small number of countries, the graduation age may shift by one due to policy changes to school duration. However, in the interest of maintaining consistency in data reconstruction, we maintain a single value of a_0 for each country defined using the most recent official school entry age and duration available. For simplicity, we omit the subscript for country and understand that the numeric value of a_0 depends on the country in question.

Ideally, the most timely observation of completion would be based on individuals one year above a_0. The reason a three age interval is considered instead is to smooth out variation resulting from the potentially small sample size of any given birth cohort in household survey data. The age bracket is shifted up by 2 years to offer a 'grace period' for delayed completion. Timely entry and progression without repetition are important goals in their own right. Nevertheless, even though children who start school late and/or repeat grades often suffer an elevated risk of drop-out, many of them *do* eventually complete school. Accordingly, the completion rate indicator, by focusing on the age group 3 to 5 years above the final grade, seeks to abstract away from the question of timeliness to some extent and capture all completion that is not unreasonably delayed.

We now define $C_{a,c,y}$ as the observed average completion at age a in a given country c in year y, such as 15-year-olds in Nigeria in 2010. As each combination of primary, lower secondary, and upper secondary school levels and female, male, and total populations are modeled independently, the level and sex indices have been omitted for simplicity. The completion rate indicator, $\mathrm{CR}_{c,y}$ is defined as a population-weighted average completion rate for individuals in the interval $[a_3, a_5]$ as follows:

$$\mathrm{CR}_{c,y} = C_{[a_3,a_5],c,y} = \sum_{i=3}^{5} \frac{p_i}{p_{[3,5]}} C_{a_i,c,y}, \tag{1}$$

where p_i is the size of the observed population aged a_i, i years above the last grade of a given level of schooling, and $p_{[3,5]}$ is the overall population in the age interval $[a_3, a_5]$.

In practice, however, population estimates and forecasts are typically based on 5-year aggregates and the single-age decompositions needed to compute $\mathrm{CR}_{c,y}$ can be subject to significant uncertainty. Additionally, an analysis of the year-to-year variation of single age cohort sizes for 10- to 14-year-olds in MICS 5 countries in Appendix A.4 suggests that random variation dominates the relative weights as opposed to smooth trends in population size. As such, for the purposes of completion rate analysis, we conclude that the population weights offer limited information but introduce extra layers of uncertainty.

Instead, for estimation purposes, we propose that an unweighted average for the completion rate indicator is preferable. While we recognise that any definitional adjustment may

cause marginal departures from observed completion rates, for the purposes of estimation and understanding trends, we use the following adjusted completion rate indicator:

$$\mathrm{CR}^{*}_{c,y} = \frac{1}{3}\sum_{i=3}^{5} C_{a_i,c,y}. \tag{2}$$

2.2 | Retrospective data

Observations of completion rates are collected from censuses and household survey programmes by aggregating individual level responses to questions on either an individual's total number of years of school completed or more directly what levels have been completed. As these nationally representative household surveys are conducted relatively infrequently, it is necessary to exploit as much information as possible from each round. If each survey only contributed estimates for the survey year for those individuals observed during the nominal age range for the completion rate indicator, many countries would have too few observations to perform any kind of robust statistical trend estimation. In the most extreme cases, there is only one survey for a country and thus one observation of individuals in the indicator age bracket.

However, censuses and household survey programmes do not just ask individuals in the $[a_3, a_5]$ age bracket about their education status, rather individuals of all ages are asked about their levels and years of school completion. Thus, one solution to the infrequent survey challenge is to take into account the education level reported as completed by older cohorts who were outside of the indicator age bracket at the time of survey.

For example, if the age bracket for the completion rate is 14–16 in a given country, then a survey in the year 2015 allows for the calculation of the 2015 completion rate based on the 14- to 16-year-olds in the sample. In addition, completion among 17- to 19-year-olds in the sample may be taken as a proxy for the completion rate among 14- to 16-year-olds 3 years prior, in 2012. More generally, $C_{a+x,c,y} \approx C_{a,c,y-x}$ for ages a that are above the expected completion age. We illustrate this process visually in Figure 2.

By leveraging observations of older age groups and tracing along cohort trajectories, a single survey contributes completion rate estimates for a series of years. This correspondence, however, is not exact. One way in which it fails is if there are many individuals completing school between the ages of a and $a + x$. Continuing the earlier example, if there are substantial delays in school progression, some 18-year-olds may complete in 2015 and thus $C_{18,c,2015} > C_{15,c,2012}$. This problem has parallels with left-censoring in that the reconstructed observations are observed after their timely completion age window has passed, thus inducing the mismatch. We illustrate this challenge in Figure 3 using observed primary school completion data from Bangladesh, Denmark and Kenya. In both Bangladesh and Kenya, observations of completion at the youngest ages are consistently lower before stabilising as the cohorts age, as would be expected in the presence of delays. Denmark, in contrast, does not have substantial delayed completion and thus the cohort trajectories reflect a near perfect correspondence.

The findings of Figure 3 suggest that in order to leverage reconstructed data for completion rate estimation, an adjustment for late completion will be necessary to ensure consistency across observations. After making such an adjustment, it is expected that the correspondence between $C_{a+x,c,y}$ and $C_{a,c,y-x}$ will be more robust but still may not be perfect. In particular, completers and non-completers may have systematically different mortality and migration rates. To limit these effects, we reconstruct completion rates only for individuals in a 20-year range starting at the bottom of the nominal age bracket (i.e. from a_3 up to a_{23}) and anchor the correspondence to a_5.

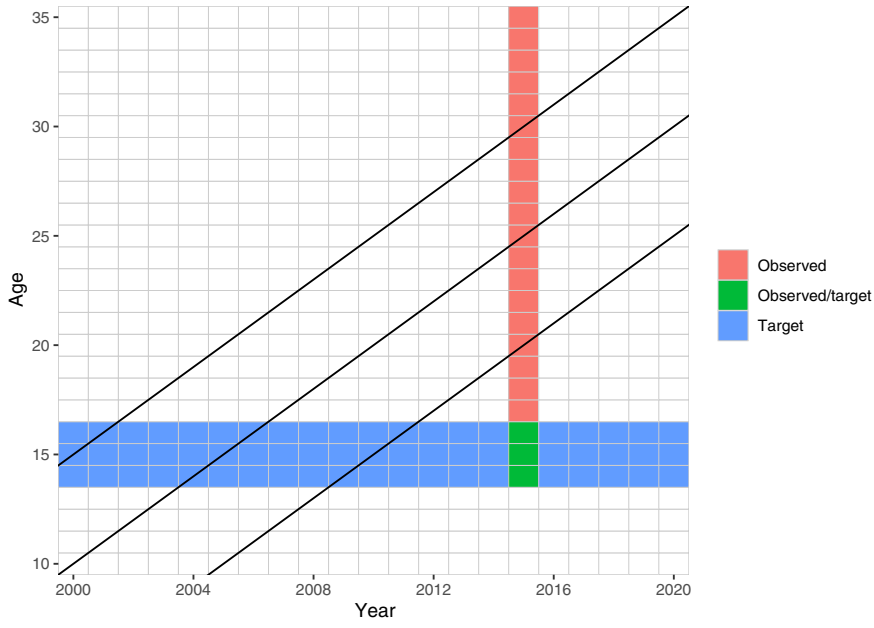

FIGURE 2 Lexis diagram for a 2015 survey with completion age bracket [14, 16]. Observed age-years are presented in red, target age bracket age-years are blue, and the overlapping age-years are green. The data reconstruction process shifts observed values into the target age bracket along the corresponding cohort trajectories. Three sample cohort trajectories are plotted with black lines. [Colour figure can be viewed at wileyonlinelibrary.com]

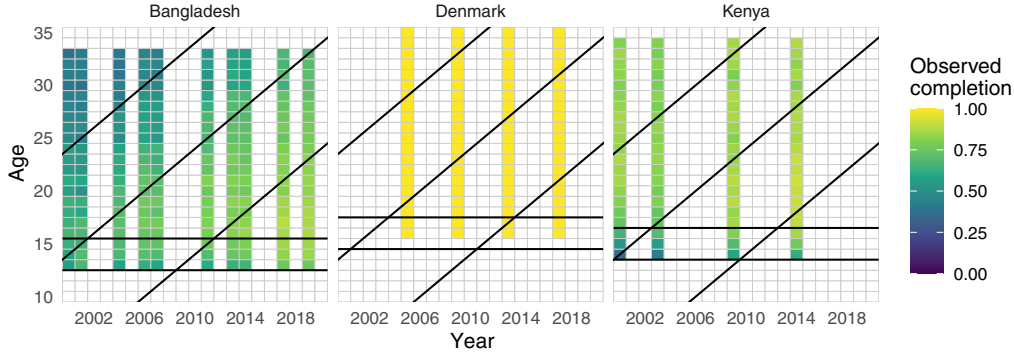

FIGURE 3 Lexis diagrams for Bangladesh, Denmark and Kenya. Observed primary completion rates are plotted according to the colour gradient in the legend. Indicator age brackets are plotted with horizontal black lines. Note that the microdata for Denmark does not include 15-year olds. Three sample cohort trajectories are plotted with diagonal black lines. [Colour figure can be viewed at wileyonlinelibrary.com]

The results of the retrospective data reconstruction process for a country with many surveys are illustrated in Figure 4. Note that the overlapping series of retrospective completion rates make it unambiguously clear that differences between surveys are often not driven by true changes in the years between the surveys, but reflect different baseline bias. That is, some surveys give systematically higher or lower estimates of completion than others.

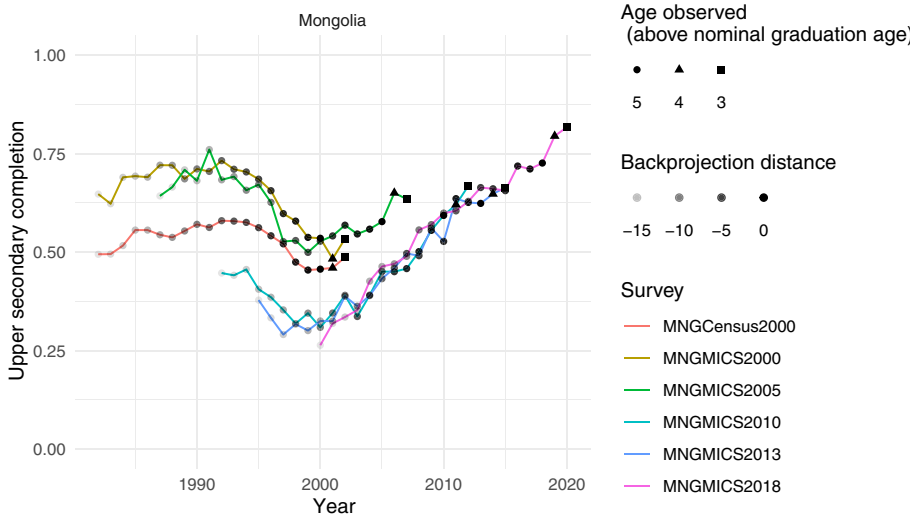

FIGURE 4 Retrospective series of completion from different surveys. Observations are aligned with the time axis according to the year in which they are at age 5 years above the nominal age for the final grade. Triangles and squares indicate observations made at younger ages. Faintness indicates retrospective observations of older cohorts. [Colour figure can be viewed at wileyonlinelibrary.com]

3 | METHODS

The objective of the ABCR model is to consolidate observations from different surveys, provide estimates for years without a survey, and allow for short-term 'now-casts' of current completion rates.

Before specifying the model, we recognise that across different levels of schooling and countries, observed completion rates cover practically the entire possible range from 0% to 100%. To account for outcomes constrained to, but spread across the entire $[0, 1]$ interval, we model observations $K_{a,c,y} = \Phi^{-1}(C_{a,c,y})$, where Φ^{-1} is the probit function.

3.1 | Model summary

We start by presenting a full completion rate model summary before describing in detail each of the component parts in the subsequent sections.

We parametrise the model in terms of the top of the age interval for the completion rate indicator, that is, we take a_5 as the reference age. Let $\kappa_{c,y} = \Phi^{-1}(\Gamma_{c,y})$ refer to the unknown *true* completion rate for the a_5 cohort in country c and year y at the outcome scale ($\Gamma_{c,y}$) and transformed scale ($\kappa_{c,y}$), respectively.

We then assume that K_i, the observed probit completion rate relating to age $a[i]$, country $c[i]$, year $y[i]$, and originating from survey $s[i]$ is distributed as follows:

$$K_i | \kappa_{c[i],y[i]}, \beta_{s[i]}, \tau_{c[i]}, \phi_{a[i],c[i]}, \nu_i, \omega_{a[i],s[i]} \sim \mathcal{N}(\kappa_{c[i],y[i]} + \beta_{s[i]} - \tau_{c[i]} \cdot \mathbb{1}_{5|a[i]} + \phi_{a[i],c[i]}, \nu_i^2 + \omega^2_{a[i],s[i]}), \tag{3}$$

where $\kappa_{c[i],y[i]}$ is the 'true' probit completion rate for the a_5 cohort in the respective country and year, $\beta_{s[i]}$ refers to the survey bias, $\tau_{c[i]}$ is a distortion due to age-misreporting occurring when $a[i]$ is a multiple of 5, $\phi_{a[i],c[i]}$ is the late (relative to a_5) completion adjustment, and finally, ν_i^2 and $\omega^2_{a[i],s[i]}$ refer to the sampling and non-sampling variances, respectively.

The model is structured in two stages. The first is the process model governing the changes in underlying true probit completion rates over time. We discuss the process model in detail in Section 3.2. The second describes through Equation (3) how the underlying true a_5 probit rates relate to the observed data. This relationship is comprised of tangible delays in completion, and various data effects including survey bias and age-misreporting. Late completion is described in Section 3.3 followed by a discussion of the data effects in Section 3.4.

After extracting the *true* a_5 completion rates, $\kappa_{c[i],y[i]}$, from the model, we can estimate the *true* completion rate indicator using the corresponding $\kappa_{c,y}$ values adjusted for within-bracket late completion as follows:

$$\widehat{CR^*}_{c,y} = \frac{1}{3}\sum_{a=a_3}^{a_5} \Gamma_{c,y} = \frac{1}{3}\sum_{a=a_3}^{a_5} \Phi(\kappa_{c,y} + \phi_{a,c}). \tag{4}$$

Also of interest is the 'ultimate cohort completion', which in our specification is assumed to be reached by 8 years after the nominal age for the final grade, at age a_8, and therefore is proxied by $\Phi(\kappa_{c,y} + \phi_{a_8,c})$.

The model in its entirety is summarised in Figure 5. This chart links all of the pieces of the model together to provide a big picture view of the completion rate estimation process.

3.2 | Core model for the underlying trend

Based on an understanding of the underlying social and policy processes determining completion rates, we wish to allow for the possibility that outcomes in a given year can have both short- and long-term repercussions. A specification in terms of first differences, that is, in *changes* in completion, better captures our intuition regarding the long-term persistence of shocks. In particular, it is reasonable as a baseline assumption that after a 'lost decade' of exceptionally poor outcomes, the average *growth* in completion will eventually return to its long-run trend. However, while it is certainly possible to make up for lost time, there is no compelling reason to think that the expected *level* of completion will eventually return to where it would have been in the absence of the crisis period.

Our core model for $\kappa_{c,y}$ is an ARIMA(1,1,0) with drift process. In addition to meeting the above requirements, the ARIMA(1,1,0) specification captures a year-over-year autocorrelation relationship, reflecting the possibility of multi-year educational development enablers or hurdles outside of the long-term drift:

$$\Delta\kappa_{c,y} = \kappa_{c,y} - \kappa_{c,y-1} = \gamma_c + \rho_c \Delta\kappa_{c,y-1} + \epsilon_{c,y}, \tag{5}$$

$$\epsilon_{c,y} | \sigma_\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{6}$$

with priors:

$$\sigma_\epsilon \sim \text{Gamma}(2, 0.1), \tag{7}$$

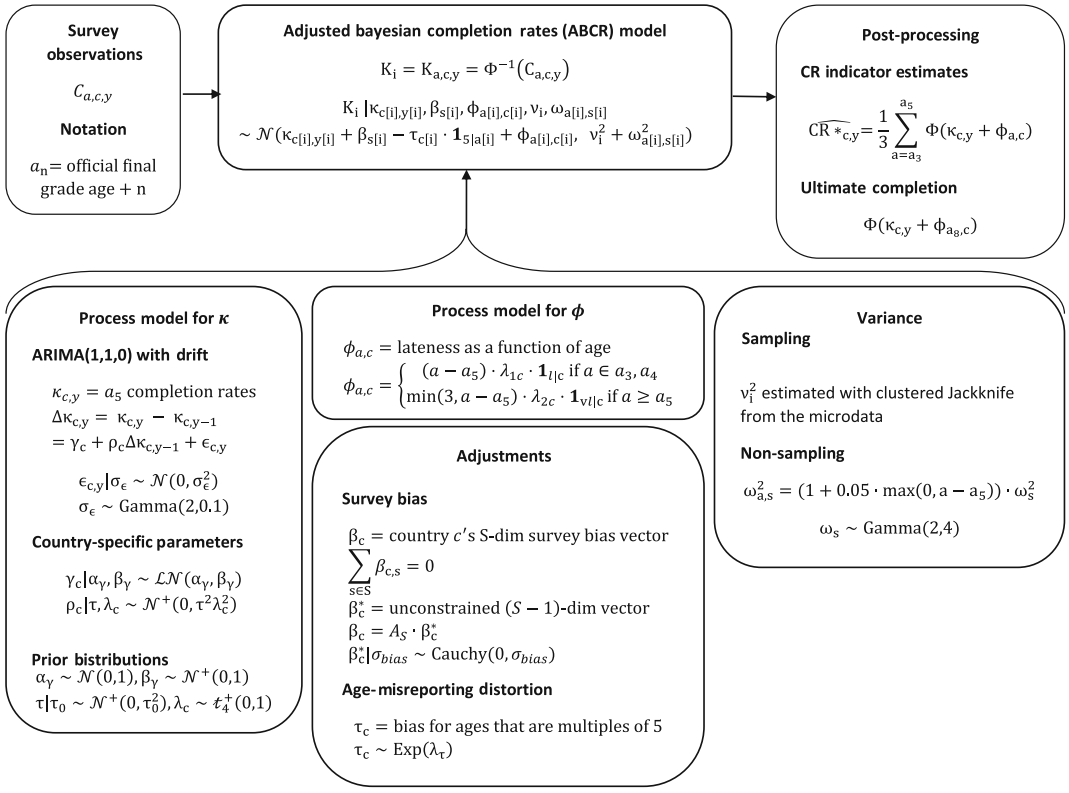

FIGURE 5 Completion rate model summary. The inputs, components and outputs of the model are summarised here and described in further detail in the following sections.

$$\rho_c | \tau, \lambda_c \sim \mathcal{N}^+(0, \tau^2 \lambda_c^2), \tag{8}$$

$$\tau | \tau_0 \sim \mathcal{N}^+(0, \tau_0^2), \tag{9}$$

$$\lambda_c \sim t_4^+(0,1) \tag{10}$$

$$\gamma_c | \alpha_\gamma, \beta_\gamma \sim \text{Log-normal}(\alpha_\gamma, \beta_\gamma), \tag{11}$$

$$\alpha_\gamma \sim \mathcal{N}(0,1), \tag{12}$$

$$\beta_\gamma \sim \mathcal{N}^+(0,1). \tag{13}$$

The long-term drift γ_c, is expected to be positive, implying an eventual convergence to 100% completion, including for upper secondary. The log-normal distribution is thus selected to reflect the positive constraint while preferring conservative drift estimates, but still allowing for the possibility of faster growth outcomes if there is strong evidence. The parameters of the log-normal distribution are assigned vague normal priors.

Some countries, however, have experienced departures from consistent growth as seen with Mongolia in Figure 4. This behaviour of short- to medium-term shocks sustained over a number of years is captured through the country-specific autoregressive coefficient, ρ_c. We note that the

autoregressive structure of an ARIMA(1,1,0) specification could model consistent growth through cascading shocks, creating a possible redundancy with the long-term drift, γ_c. Given that there is, however, substantive interest in understanding the differences in strength of consistent growth between countries (Malala Fund, 2014; United Nations, 2015), the specification should prefer to model consistent growth with γ_c instead of with ρ_c. To reflect that ρ_c is only to be relevant in the presence of medium-term shocks that are only experienced by a subset of countries, we assign ρ_c a horseshoe prior to shrink unneeded ρ_c terms towards zero (Carvalho et al., 2009). Assuming a prior assumption that 30% of countries will have a relevant ρ_c term and a $t_4^+(0,1)$ distribution on the country-specific horseshoe parameter, λ_c, we set the hyperparameter $\tau_0 = 0.01$ following the procedure discussed in Piironen and Vehtari (2017). Appendix A.1 provides additional comments on both the horseshoe specification and the process model as a whole.

Given that the model is structured in two stages, we employ a vague boundary avoiding prior (Chung et al., 2013; Stan Development Team, 2020b) for σ_ϵ to discourage the scenario where the underlying κ process exhibits little to no variability. Since the true completion rates are unobserved, a value of σ_ϵ approaching zero could be consistent with the likelihood if the variance of the data generating model in Equation (3) expands to entirely compensate for the variability in observations.

3.3 | Late completion

In a number of countries where delays in school entry and progression are severe, even the 'grace period' allowed by the shifted age bracket is not sufficient to ensure that $C_{a_3,c,y}$ already represents ultimate completion of the cohort in question and equals $C_{a_5,c,y+2}$. In other words, some individuals complete school *during* the age interval $[a_3, a_5]$ and, in some cases, even beyond. This is clearly evident in the example in Figure 6. Observations at ages a_3 and a_4 consistently display lower completion than observations at age a_5, and observations at a_5 consistently display lower completion than observations at ages a_6 through a_8.

For a single cross-sectional age profile of completion from one survey, such a pattern would not establish late completion but could, in principle, also arise from a decline in completion between successive cohorts. However, overlaying the retrospective completion rates from several surveys, as in Figure 6, amounts to an implicit pseudo-cohort analysis that shows that ultimate completion suffered no such decline. Instead, the completion observed in a given survey for some pseudo-cohort depends on the age at which it is observed, even at ages above a_3. Indeed, in Figure 6, it is evident that late completion continues even past a_5 in Liberia and Malawi.

As discussed in Section 2.2, to ensure that the retrospective data are consistent and comparable, it is therefore necessary to model the age profile of observed completion, in a way that allows for late completion in addition to the error associated with retrospective observations. As a parsimonious but flexible specification, we model the late completion effect $\phi_{a,c}$ as a piece-wise linear function in the probit space with two segments as follows:

$$\phi_{a,c} = \begin{cases} (a - a_5) \cdot \lambda_{1c} \cdot \mathbb{1}_{l|c} & \text{if } a \in \{a_3, a_4\} \\ \min(3, a - a_5) \cdot \lambda_{2c} \cdot \mathbb{1}_{vl|c} & \text{if } a \geq a_5. \end{cases}$$

The first case specifies completion within the $[a_3, a_5]$ interval potentially being lower by a country-specific value λ_{1c} per year. The second case models additional very late completion

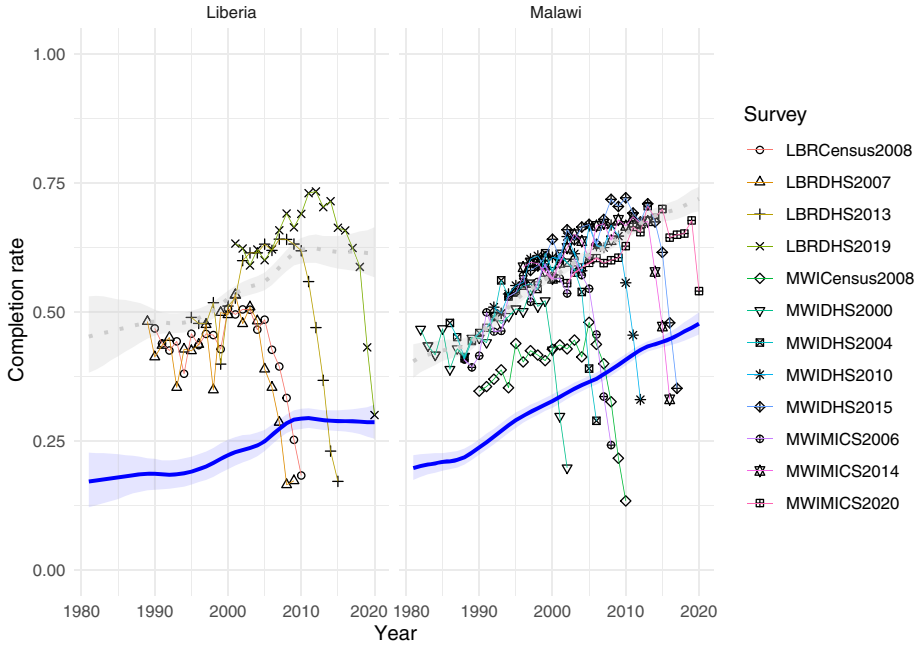

FIGURE 6 Observed and backcast values of age-specific primary completion in Liberia and Malawi. Observations corresponding to a_3, a_4 and a_5 for each survey are the first, second and third points from the right of each data series, respectively. The fitted completion rate indicator and ultimate completion are indicated by the blue and grey lines, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

beyond the CR indicator interval with a country-specific slope, λ_{2c}, accumulating over three years to the age of 'ultimate cohort completion', a_8. The choice of three additional years of delay is based on an assessment of surveys from countries in Sub-Saharan Africa, the region with the most significant delays. In these countries, peak completion is typically achieved by a_8 at the latest suggesting a total of 8 years of delays is appropriate. This assumption is further validated in Section 4.2 where we study the sensitivity of the model to different very late completion durations. Visually, the age specification is described by Figure 7.

Both late completion effects are subject to indicator variables dictating the presence or absence of the effect for the given country. Briefly, highly developed education systems have limited or negligible delayed completion beyond the three year grace period offered by the indicator age bracket. To reflect this fact, those countries with median observed completion rates above 0.95 are assumed to not have structural late completion and instead any minor dips are considered noise. However, if for a country with close-to-universal completion, the observations for $[a_3, a_5]$ are consistently below those of $[a_5, a_7]$ across surveys, shorter-term late completion (i.e. $\lambda_1 c$) is estimated. Additional details of the procedure for identifying the presence of late completion are provided in Appendix A.2. The country-specific parameters, λ_{1c} and λ_{2c} are modelled hierarchically with the following priors:

$$\lambda_{1c}|\sigma_{\lambda_1} \sim \mathcal{N}^+(0, \sigma^2_{\lambda_1}). \tag{14}$$

$$\lambda_{2c}|\sigma_{\lambda_2} \sim \mathcal{N}^+(0, \sigma^2_{\lambda_2}). \tag{15}$$

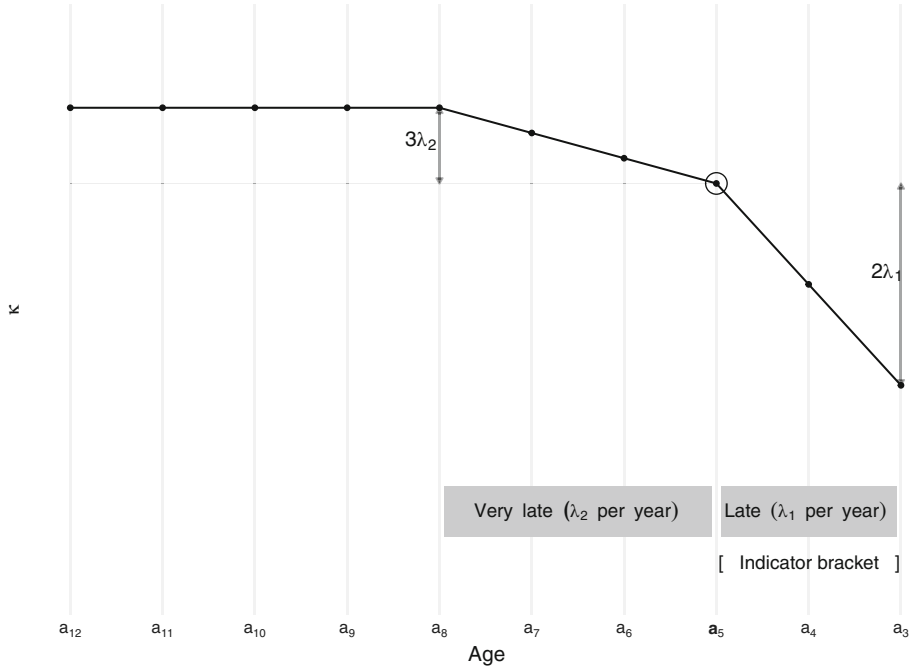

FIGURE 7 Piece-wise linear model of age-profile of completion. Note that age is decreasing on the *x*-axis to match a retrospective data series.

$$\sigma_{\lambda_1} \sim \mathcal{N}^{+}(0,1). \tag{16}$$

$$\sigma_{\lambda_2} \sim \mathcal{N}^{+}(0,1). \tag{17}$$

3.4 | Data considerations

In addition to the real year-over-year changes in completion and delays in completion, there are consequences of using household surveys that must be accounted for. The three components of these data considerations are survey bias, age-misreporting, and differences in variability.

3.4.1 | Survey bias

In the present case of completion rates, no equivalent to the— theoretical—'gold standard' of a complete vital registration system for health applications or specialised in-depth studies for MMR exists. Even censuses may miss important subgroups, such as street children. This issue is particularly consequential for completion rates as differences in completion between included and excluded groups are potentially extreme. It is entirely plausible for primary completion to be almost universal among population in households, but close to zero among 'missing children'. Given the lack of a 'gold standard', *absolute* survey bias cannot be modelled without introducing strong assumptions. However, modelling *relative* bias allows the model to understand systematic differences between surveys even in periods where retrospective series constructed from different

surveys do not overlap. This relative structure induces a sum-to-zero constraint on each country's collection of survey bias terms. Thus, to model β_c, the vector of country c's S survey bias terms, β_s, we parameterise in terms of β_c^*, a vector of $S-1$ elements and assume the following:

$$\beta_c = A_S \cdot \beta_c^*, \tag{18}$$

$$A_S = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\frac{1}{S-1} & 1-\frac{1}{S-1} & 0 & \dots & 0 \\ -\frac{1}{S-1} & -\frac{1}{S-1} & 1-\frac{2}{S-1} & & 0 \\ \vdots & \vdots & & \ddots & \\ -\frac{1}{S-1} & -\frac{1}{S-1} & \dots & -\frac{1}{S-1} & 1-\frac{S-2}{S-1} \\ -\frac{1}{S-1} & -\frac{1}{S-1} & -\frac{1}{S-1} & \dots & -\frac{1}{S-1} \end{bmatrix}, \tag{19}$$

$$\beta_c^* | \sigma_{\text{bias}} \sim \text{Cauchy}(0, \sigma_{\text{bias}}), \tag{20}$$

$$\sigma_{\text{bias}} \sim \mathcal{N}^+(0, 0.25^2), \tag{21}$$

where A_S is an $S \times (S-1)$ matrix with elements a_{ij} constrained such that $\forall j \in 1, \dots, S-1$, $\sum_{i=1}^{S} a_{ij} = 0$, and $\forall i \in 1, \dots, S$, $\sum_{j=1}^{S-1} |a_{ij}| = 1$. The first constraint enforces the sum-to-zero constraint using the columns of A_S, and the second constraint propagates the symmetric Cauchy$(0, \sigma_{\text{bias}})$ prior from β_c^* to β_c. The Cauchy distribution is selected here to address the possibility of potentially extreme bias in specific surveys. Additional details, are provided in Appendix A.3.1.

3.4.2 | Age-misreporting distortion

It is well known that in developing country settings, respondents' ages may be misreported, leading to an over-representation of ages that are multiples of five. It is also known that this behaviour correlates with low numeracy skills. Accordingly, it is plausible that cohorts with ages that are multiples of five at the time of a survey may have understated completion rates due to the over-representation of individuals with low numeracy skills that have misreported their ages. When reconstructing retrospective data, this would manifest as anomalous drops in completion every 5 years.

Indeed, we see clear evidence of this in Figure 8, for example. Here, as in a number of other cases, reported primary school completion is lower among those whose reported age is a multiple of five. This is what would be observed if those who did not complete primary school are more likely to round their age.

Retrospective observations that represent a reported 'round' age group at the time of survey are coded with an indicator variable. Observations where this indicator equal 1 are subject to an additional term τ_c in Equation (3) that accounts for the potential distortion in country c due to age-misreporting. This distortion is parsimoniously modelled as being rare, but potentially large with the following hierarchical structure:

$$\tau_c | \lambda_\tau \sim Exp(\lambda_\tau). \tag{22}$$

$$\lambda_\tau \sim \mathcal{N}^+(0, 50^2). \tag{23}$$

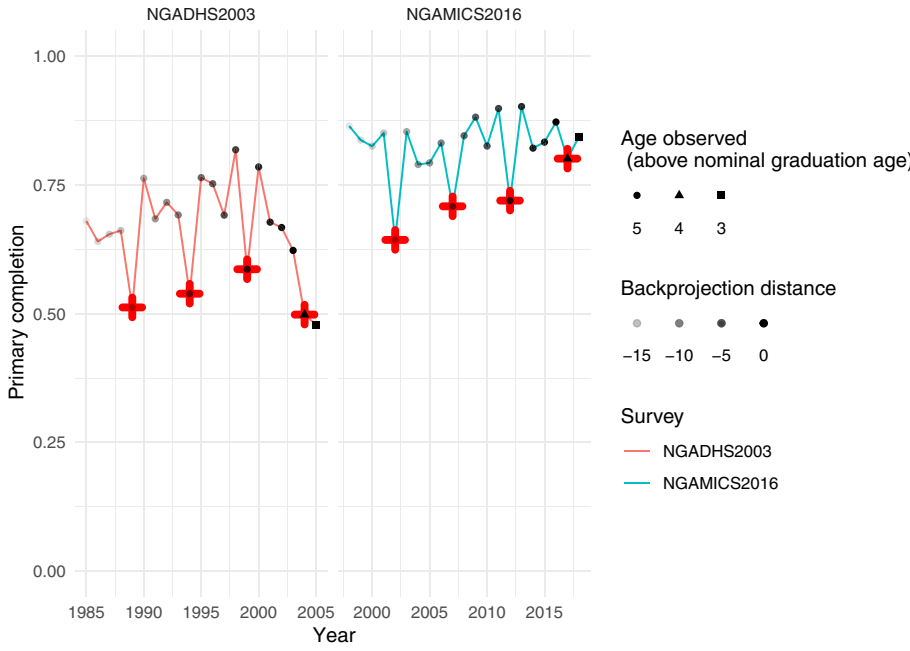

FIGURE 8 Observed and backcast values of age-specific primary completion in Nigeria. Red plus signs indicate observations based on respondents reporting their age at time of survey as a multiple of five. [Colour figure can be viewed at wileyonlinelibrary.com]

We note that the specification for age-misreporting is simple. A more complex model could perhaps link the magnitude of age-misreporting to the pool of less-educated likely to misreport. For example, if the prevalence of age-misreporting were a function of primary completion, the estimates of primary completion could be incorporated into the specification. However, such a step would mean that the estimation of completion rates at each level could no longer be done independently. Too little is known to do this convincingly and so we settle on the simple specification above. Similarly, in some cases it seems as if the adjacent 'almost round' ages report increased primary school completion as a result of losing some of their unschooled who incorrectly place themselves in the round age group. However, in other cases the offsetting increase is more diffuse. Given the lack of consistent pattern across surveys and countries, the offset is not modelled explicitly as affecting specific ages, but is allowed to be implicitly absorbed in the overall country intercept.

3.4.3 | Variance

To account for differences in observation specific sampling variation between surveys (and between age groups), sampling variances, v_i, are estimated a priori and provided as input. As survey reports do not provide sampling errors for completion rates, v_i is estimated from the micro-data, applying the clustered Jackknife procedure used to generate the published DHS SE estimates for other indicators (The Demographic and Health Surveys Program, 2012).

We also assume that the observation for age a from a specific survey s is subject to an independent *non-sampling* error, $\omega_{a[i],s[i]}$. We noted previously that retrospective data may face increasing uncertainty as time passes due to different mortality and migration rates between completers

and non-completers. By restricting the age range for reconstruction to $[a_3, a_{23}]$, the effect of mortality is mostly limited outside of jurisdictions with high young adult mortality rates stemming from high HIV/AIDS prevalence. In general, differential migration between completers and non-completers is expected to have a larger impact within the age range. To reflect the increasing uncertainty as the retrospective distance increases, we first estimate an underlying survey level value ω_s and then scale it linearly with the retrospective estimation distance to reflect the increased uncertainty as time passes. Specifically, $\omega^2_{a,s} = (1 + 0.05 \cdot \max(0, a - a_5)) \cdot \omega^2_s$ with $\omega_s \sim$ Gamma(2, 4). The gamma prior is selected for its boundary avoiding properties as a zero value for non-sampling variance is inconsistent with the understanding that non-sampling variance is present in this context. In principle, if not only the migration intensity, but also the migration age schedule differs between completers and non-completers, the magnitude of the retrospective estimation error could be a non-linear function of the elapsed time x or equivalently, of age at the time of survey. In practice, in the absence of a priori information on this effect, we assume that the uncertainty associated with retrospective estimation increases linearly with age such that the uncertainty doubles over a 20-year interval. Further details are presented in Appendix A.3.3.

3.5 | Implementation

The model parameters are estimated in a Bayesian framework using R. Samples from the posterior distribution are generated using the No-U-turn sampling (NUTS) Hamiltonian Monte Carlo algorithm (Hoffman & Gelman, 2014; Neal, 2011) implemented in the Stan package (Carpenter et al., 2017; Stan Development Team, 2020c). Four chains are run in parallel for each level and sex combination. Each chain consists of 3000 burn-in iterations, and 3000 samples. Due to memory constraints, the final sample is thinned to 1000 per level and sex combination. We check convergence using the standard diagnostic checks including trace plots, pairs plots and the Gelman and Rubin diagnostic (Gelman & Rubin, 1992; Stan Development Team, 2020a; Vehtari, Gelman, Simpson, et al., 2020). See Appendix B for additional implementation details.

3.6 | Validation

To assess the performance of the model, we consider two different out-of-sample validation exercises. First, we conduct a 'leave one survey out' validation. Specifically, all observations based on the latest survey (including backcast values) for each country with more than one survey are omitted from the estimation of the models, and predicted values for these values are obtained. Such a validation is designed to mimic an intended use case of the model, that is, comparing the output to a new survey. As entire surveys are left out, no survey bias is estimated for the test surveys and so we compare the left out values to the appropriate $\kappa_{c,y}$, adjusted for age-misreporting and late completion, by computing mean squared errors (MSE) and mean absolute errors (MAE).

The second exercise leaves out two random observations from each survey, provided each survey has at least five observations. Notably, this does not completely remove any surveys from the data and so bias terms for all surveys can be computed. For this test, we also compute the MSE and MAE between the left out observations and the respective $\kappa_{c,y}$ adjusted for age-misreporting, late completion, and survey bias. We also compute model bias, and coverage of prediction intervals defined as $n^{-1}\sum_{i=1}^{n} \mathbb{1}_{l_i \le y_i \le u_i}$ where n is the number of observations in the test set, i is the current observation, and l_i and u_i are the lower and upper bounds, respectively, of the prediction interval.

4 | RESULTS

The model output is illustrated by the country-level results shown in Figure 9 for a selection of countries at three different levels of schooling. The presented examples have been selected to illustrate a variety of scenarios. The results appear sensible, capturing late completion where appropriate and with projected uncertainty greater when fewer or even only a single survey was available, for instance. Particular attention should be given to the Nigeria, Rwanda, Bhutan and Armenia examples.

As the country with the single most surveys, the observations for Nigeria appear incredibly noisy. However, as observed in Figure 8, much of this 'noise' is actually the manifestation of age-misreporting. The ABCR model considers this information and produces estimates with greater certainty than the underlying data might suggest at face value. Rwanda presents an example of the model's ability to adapt to non-standard patterns. While the long-term trend is positive, the effects of the 1994 genocide are clearly visible as a large drop in completion observed and captured by the model. Bhutan presents a single survey case. Given the single survey presents remarkable improvements in completion, a strong positive trend is produced. This is offset, however, by significant uncertainty throughout the series. Finally, Armenia demonstrates the behaviour when one or two surveys greatly deviate from the rest, whether as a result of design or implementation problems, or inconsistent coding during analysis. The divergent surveys do contribute information to the estimated *true* completion rates, but do not dramatically pull down the estimates. The Cauchy distribution for survey bias permits large survey biases in rare instances such as this that would be irreconcilable if an alternative such as a Normal distribution had been selected.

4.1 | Fit

In the present case, it is not clear what benchmark the ABCR model should be compared to by default as there is no previous attempt at modelling the same outcomes using a simpler specification. However, basic linear models are used in practice to project SDG 4 indicators (EFA Global Monitoring Report, 2015; UNESCO Institute for Statistics, Global Education Monitoring Report Team, 2019). We thus consider a 'simple' Bayesian model in which the κ for a given country is a linear function of an intercept and slope over time. In other word, it fits a plain probit curve to the C_i, without taking into account differences in sampling variation, survey bias, common parameter distributions or any of the advanced data considerations specified in the ABCR model. Additionally, we compare the ABCR model against reduced versions of itself. In the 'ABCR (No Late)' model, we drop the late completion term ϕ, in the 'ABCR (No Bias)' model, we drop the survey bias term β, and in the 'ABCR (No Distortion)', we drop the age-misreporting distortion term τ. Examining the performance of these reduced models offers insight into the contribution of each of the adjustment components to the model.

Results from the 'leave one survey out' validation exercise for the total population are presented in Table 1. Results for the separated male and female populations can be found in Appendix A.5. The MSE and MAE of all five models are presented multiplied by 100 for ease of reading. Recall that in this setup, the bias of the target surveys are not exploited, since empirically, the survey series is not informative of the bias to be expected of a given individual survey (see Figure 12).

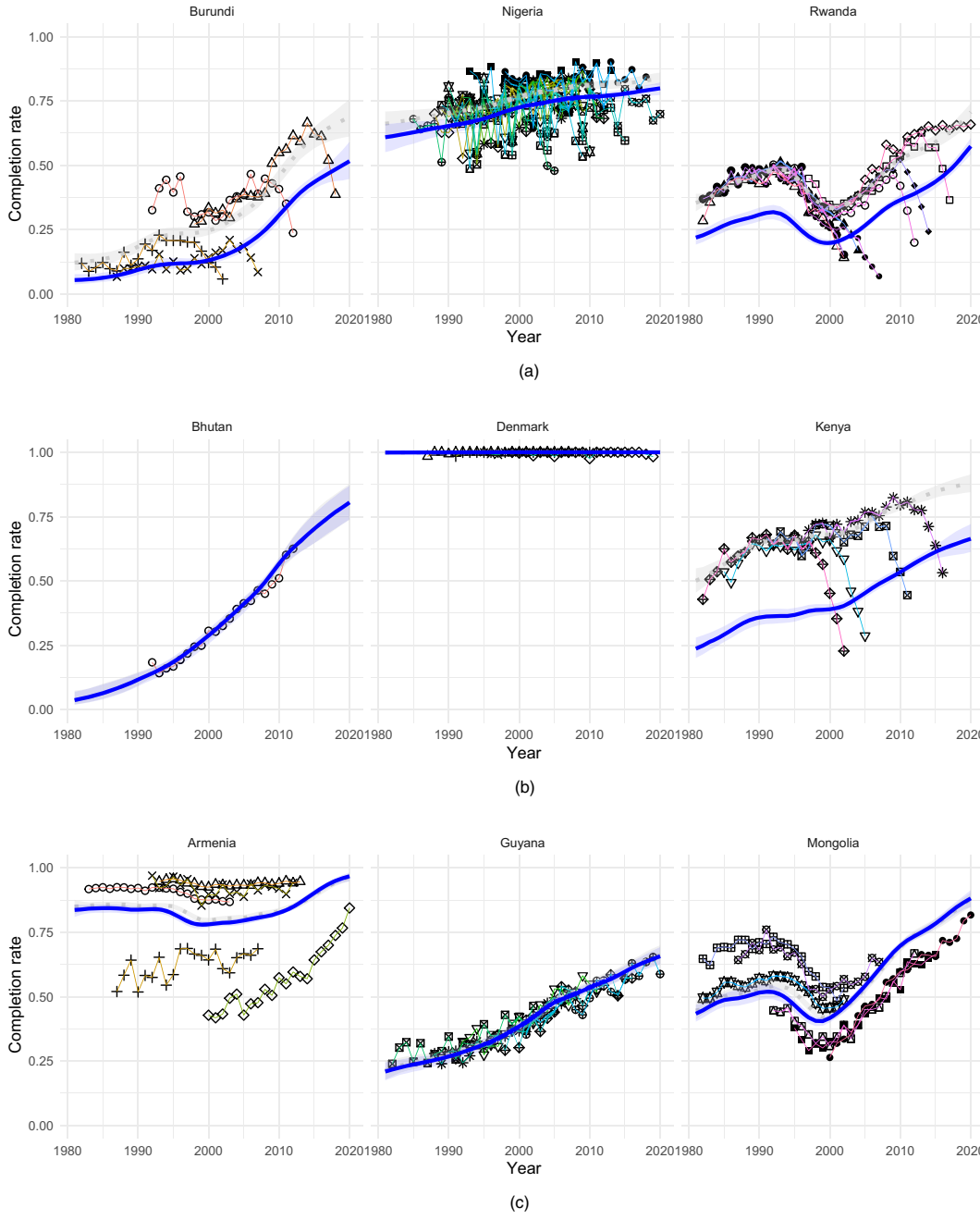

FIGURE 9 Country projection by level. The solid blue and dashed grey lines denote the completion rate indicator and ultimate completion, respectively. Observations originating from each individual survey are distinguished by a distinct shape and colour combination. (a) Primary; (b) lower secondary; (c) upper secondary [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Adjusted Bayesian completion rates (ABCR) 'leave one survey out' validation results

Error (×100)	ABCR	Simple	ABCR (no late)	ABCR (no bias)	ABCR (no distortion)
Primary					
MSE	0.26	0.44	0.87	0.23	0.28
MAE	3.00	3.91	5.01	2.83	3.08
Lower secondary					
MSE	0.53	0.89	0.91	0.51	0.53
MAE	4.50	5.74	5.94	4.41	4.50
Upper secondary					
MSE	0.88	1.01	1.12	0.96	0.89
MAE	6.32	6.89	7.43	6.45	6.37

The ABCR model offers a meaningful and worthwhile improvement on the simple specification across education levels and populations. On the whole, the advantage of the ABCR model is higher for lower levels of schooling where effects such as late completion appear more pronounced. Similarly, the model tends to perform better on female and total population data.

The results for the reduced models corroborate these observations. Removing the late completion correction greatly reduces the value of the specification, most dramatically in the primary level. Removing the age distortion parameter is less consequential though that is expected as it only contributes to 20% of the data points and is immaterial is developed countries. The specification without bias performs comparably to the complete ABCR specification though that is expected in this test given that survey bias is not included in predictions due to the complete removal of surveys.

The results from the random out-of-sample exercise are presented in Table 2. The bias values suggest the model does not experience significant bias. In this validation exercise, the complete ABCR specification consistently outperforms the simple and reduced models. Similar to the previous exercise, the removal of the late completion parameter materially worsens the results whereas the removal of the age-misreporting distortion has milder consequences. Now that survey bias is exploited in predictions, it is clear that the survey bias term greatly reduces the model's errors.

In addition to computing errors and bias in the random out-of-sample exercise, we also compute coverage values in Table 3. Across specifications, the reported coverage values are in general close to the nominal level though there are slightly fewer out-of-sample observations outside of the prediction intervals than expected. We note that the tendency to produce conservative predictions is consistent with the in-sample findings in Appendix B.4 and is preferred over the contrary.

The out-of-sample validation exercises performed suggest that the model outperforms the simple specification and is fairly well calibrated. When comparing the results of the complete ABCR specification against the reduced models, it is clear that each of the late completion, survey bias, and age-misreporting distortion terms contribute to the model's performance. Of the three terms, late completion and survey bias appear to be have the greatest contributions. Additional supporting in-sample visuals and commentary are provided in Appendix B.4.

TABLE 2 Adjusted Bayesian completion rates (ABCR) random test set validation results

Error (×100)	ABCR	Simple	ABCR (no late)	ABCR (no bias)	ABCR (no distortion)
Primary					
MSE	0.07	0.33	0.18	0.14	0.09
MAE	1.52	3.19	2.25	2.11	1.71
Bias	0.11	0.25	0.18	−0.03	0.13
Lower secondary					
MSE	0.08	0.39	0.13	0.26	0.10
MAE	1.84	4.03	2.25	2.96	2.05
Bias	0.16	0.35	0.22	0.00	0.13
Upper secondary					
MSE	0.09	0.39	0.11	0.41	0.10
MAE	2.07	4.22	2.27	3.59	2.18
Bias	−0.03	−0.15	−0.04	−0.74	−0.07

TABLE 3 Adjusted Bayesian completion rates (ABCR) random test set coverage results

Coverage level (%)	ABCR	Simple	ABCR (no late)	ABCR (no bias)	ABCR (no distortion)
Primary					
0.80	0.89	0.93	0.90	0.88	0.89
0.90	0.95	0.97	0.96	0.95	0.96
0.95	0.98	0.98	0.97	0.98	0.98
Lower secondary					
0.80	0.87	0.90	0.88	0.85	0.87
0.90	0.94	0.95	0.94	0.94	0.94
0.95	0.97	0.97	0.96	0.97	0.97
Upper secondary					
0.80	0.85	0.90	0.85	0.86	0.86
0.90	0.95	0.96	0.94	0.95	0.94
0.95	0.98	0.98	0.97	0.98	0.98

4.2 | Sensitivity

In the specification of the three adjustment factors, late completion, survey bias and age-misreporting, we assign hyperpriors to parameter prior distributions to avoid imposing highly rigid assumptions. Elsewhere, there are three key assumptions in the model where we instead perform a sensitivity analysis to study how the model responds. Specifically, we study changing the prior distribution assigned to the non-sampling variance, ω from Gamma$(2, 4)$ to Normal$(0, 1)$,

TABLE 4 Adjusted Bayesian completion rates (ABCR) sensitivity results

Error (× 100)	ABCR	Late (+0 years)	Late (+5 years)	Var ($N(0,1)$)	Var ($G(2,2)$)	Var ($G(2,8)$)	Tau0 (0.5×)	Tau0 (2×)
Primary								
MSE	0.07	0.08	0.10	0.06	0.06	0.06	0.06	0.06
MAE	1.52	1.66	1.82	1.49	1.50	1.49	1.50	1.49
Bias	0.11	−0.07	−0.81	−0.04	−0.04	−0.04	−0.03	−0.04
Lower secondary								
MSE	0.08	0.09	0.09	0.08	0.08	0.08	0.08	0.08
MAE	1.84	1.92	2.00	1.84	1.85	1.84	1.86	1.84
Bias	0.16	0.13	−0.42	0.14	0.14	0.14	0.14	0.14
Upper secondary								
MSE	0.09	0.10	0.11	0.10	0.10	0.10	0.10	0.10
MAE	2.07	2.18	2.22	2.16	2.16	2.15	2.15	2.15
Bias	−0.03	0.09	−0.39	0.09	0.10	0.09	0.10	0.10

TABLE 5 Adjusted Bayesian completion rates (ABCR) sensitivity coverage results

Coverage Level (%)	ABCR	Late (+0 years)	Late (+5 years)	Var ($N(0,1)$)	Var ($G(2,2)$)	Var ($G(2,8)$)	Tau0 (0.5×)	Tau0 (2×)
Primary								
0.80	0.89	0.89	0.84	0.87	0.90	0.89	0.89	0.89
0.90	0.95	0.95	0.90	0.95	0.96	0.95	0.96	0.96
0.95	0.98	0.97	0.95	0.97	0.98	0.97	0.98	0.98
Lower secondary								
0.80	0.87	0.87	0.85	0.86	0.87	0.87	0.88	0.87
0.90	0.94	0.93	0.92	0.93	0.94	0.94	0.94	0.94
0.95	0.97	0.97	0.96	0.96	0.97	0.97	0.97	0.97
Upper secondary								
0.80	0.85	0.85	0.83	0.83	0.85	0.84	0.85	0.85
0.90	0.95	0.93	0.92	0.92	0.94	0.93	0.94	0.93
0.95	0.98	0.96	0.96	0.96	0.97	0.96	0.97	0.97

Gamma(2, 2), and Gamma(2, 8), the length of the extended late completion duration in ϕ from 3 to 0 years (i.e. ultimate completion achieved at a_5) and 5 years (i.e. ultimate completion achieved at a_{10}), and the prior value of τ_0 from 0.01 to 0.005 and 0.02 in the ρ specification. We perform an out-of-sample exercise using the total population data for each education level where two random observations from survey are removed, provided each survey has at least five observations. Table 4 provides the computed MSE, MAE and bias values, and Table 5 provided coverage results.

The model appears robust to changes in the non-sampling variance and the initial value of τ_0 given the consistency in the errors, bias and coverage. The duration of the extended late

completion assumption is more interesting. Both shortening and lengthening the period from the default 3-year period worsen the results slightly. This is perhaps expected as eliminating the extended delay is inadequate for countries like Liberia and Malawi in Figure 6, yet there is also no substantial evidence to suggest that more than a 3-year extension is needed.

4.3 | Posterior parameter estimates

In addition to the posterior estimates of the outcome, we examine the estimates for specific model parameters relating to phenomena of substantive interest. Figure 10 summarises the distributions of the median estimates for survey bias, late completion and age-misreporting.

Recall that the two-stage specification of the age-misreporting effect: whether it occurs, and if, then how strongly. Figure 10 displays the latter, the strength of the age-misreporting effect as it applies to the relevant observations. That the effect is of similar magnitude across education levels is not entirely surprising, because the age reporting of those without any schooling (putatively the group most likely to misreport) will distort the denominator of primary, lower, and upper secondary completion equally. In terms of its magnitude, it is clear that this effect cannot be neglected. It can exceed the typical magnitude of bias of individual surveys, and is comparable in magnitude to the distortion arising from late secondary completion.

Turning to the degree of late completion, we can see that there is large variation between countries, but largely around fairly high levels. Values in Figure 10 are shown on the scale of κ.

FIGURE 10 Distribution of posterior medians of survey bias, late completion and age-misreporting, probit scale [Colour figure can be viewed at wileyonlinelibrary.com]

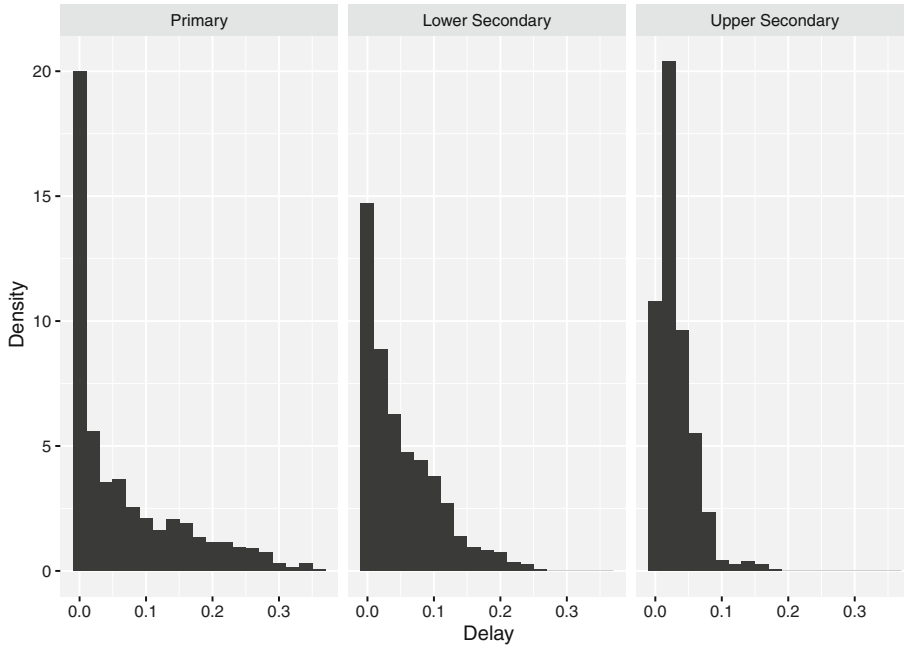

FIGURE 11 Distribution of late completion delays with the indicator age bracket on the real scale

Because of the floor and ceiling effect at 0% and 100%, respectively, the translation of the lateness parameter to the original completion rate scale depends on the level of completion it modifies due to the nonlinearity of the probit function.

To illustrate the estimated impact of late completion in the real space, in Figure 11 we plot histograms of the percentage difference in completion between the top and bottom of the age bracket for the combined female and male populations across all country-years. That is, we plot $\Phi(\kappa_{c,y}) - \Phi(\kappa_{c,y} + \phi_{a_3,c})$. In many countries, the late completion effect is small, but delays of 10% or more are prevalent in the primary and lower secondary estimates. Extreme delays of 20% or more are found in 26 countries, suggesting that major delays in schooling are deeply embedded in those education systems.

The lower level of late completion at the upper secondary level in Figure 11 is worth noting. While a longer school career could create more opportunities for repetition, our results emphatically suggest that instead of completing late, vulnerable students at higher levels, including late completers at lower levels, drop out—or are pushed out.

Because the model can only estimate *relative* survey bias, the estimates for the survey bias terms centre on zero. The presence of heavily biassed surveys was expected and encoded through the Cauchy distribution but the high frequency of extreme bias illustrated by Figure 10 is still surprising. This calls for great caution in interpreting any survey-based education indicators in countries where only a single survey has been conducted.

Figure 12 displays the uncertainty around individual bias estimates, by survey series. Note that this plot includes only the survey bias estimates from the combined male and female population models and groups country-specific non-standard surveys under the 'Other' category. We see that the conclusion that some individual surveys suffer heavy bias is confirmed with great confidence. However, it is also apparent that apart from MICS6 which displays fairly consistent positive bias, there is no discernible systematic difference between DHS and MICS series, or specific waves: the

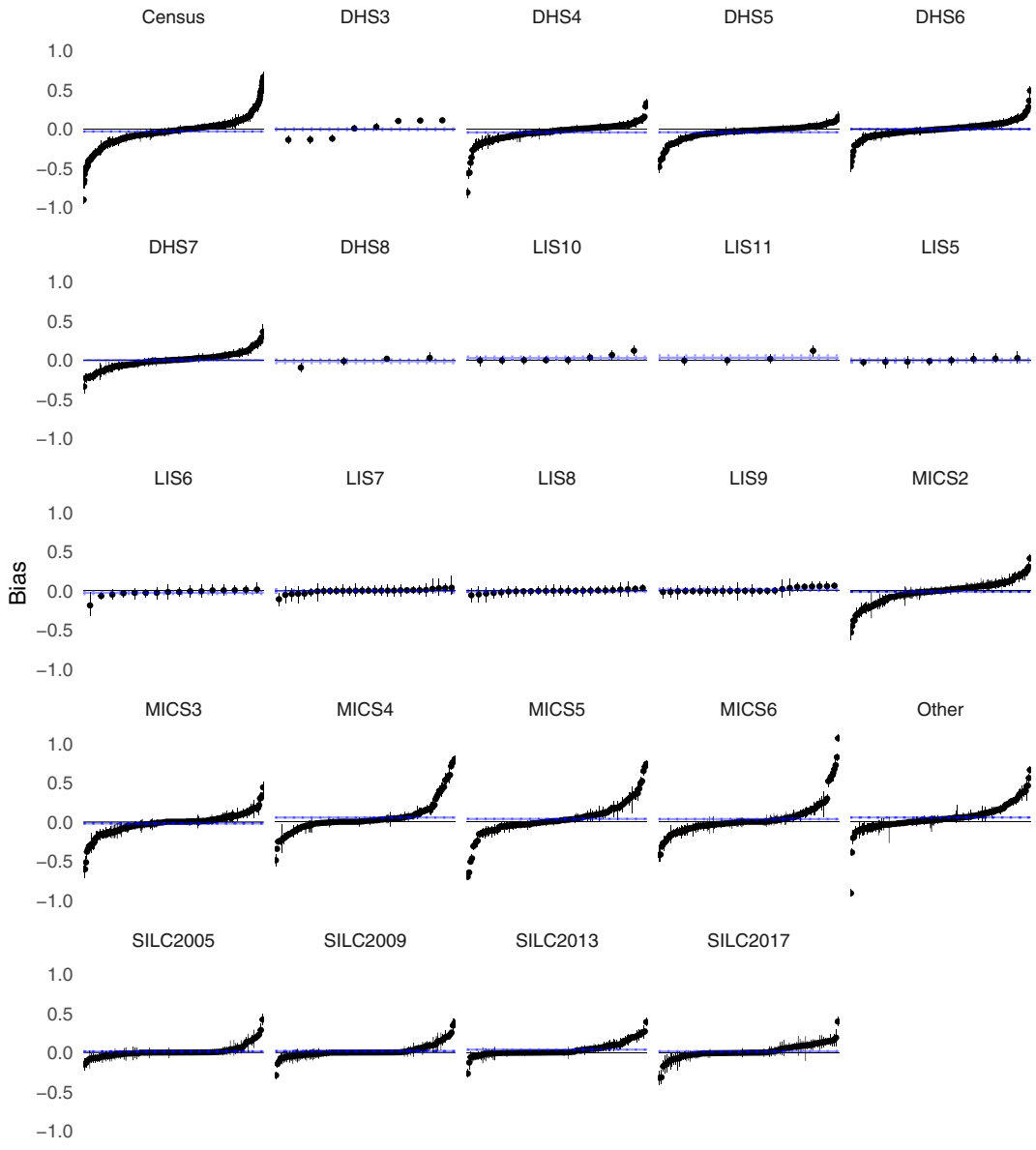

FIGURE 12 Posterior distribution of individual survey bias terms, probit scale. Blue lines: median (solid), 10th and 90th percentile (dotted) [Colour figure can be viewed at wileyonlinelibrary.com]

estimated systematic bias common to surveys of a given type is practically indistinguishable from zero. We also notice that censuses do not seem to gravitate towards lower bias values or provide any indication of being a 'gold standard'.

It is possible that the 'systematic bias' of household survey estimates of school completion is largely shared across survey designs. In other words, to the extent that some low-education groups are missing from sampling frames, such as nomads or children in orphanages, they tend to be missed by household surveys in general. If this is the case, the implication is that 'better' household surveys may not be sufficient to capture invisible groups, and that altogether, alternative approaches to complement them may be needed.

5 | CONCLUSION

In this work, we present a novel method to estimate SDG global indicator 4.1.2: completion rate in primary, lower secondary and upper secondary education. This method, the ABCR model, consolidates reconstructed survey data and addresses the significant data challenges of age-misreporting and late completion in order to estimate the underlying true completion rate. Such steps are necessary given that the ABCR model represents the first attempt at modelling children and youth completion, a 'flow' measure of education attainment, as opposed to a 'stock' measure of education attainment among adults. Comparing the model with a simpler base case validates the presented specification. The ABCR model is now used by the United Nations to estimate SDG indicator 4.1.2 and monitor regional and global progress.

We recognise that global model-based estimates of development indicators should not be mistaken for the actual measure. This is especially true when estimates are used to assess the achievement of specific time-bound targets. In the present case, at the time of writing, the most recent surveys (for a handful of countries) collected data in 2018 and 2019. The median most recent survey year across countries is 2017. We do believe our estimates for 2019, say, to be the 'best guess' for the situation given past dynamics.

In the immediate future, however, there is much uncertainty regarding the effects the COVID-19 pandemic will have on education. We believe that our model's current short-term projections can act as a useful baseline with which to answer the 'what could have been' question regarding the impact of COVID-19. Further, as new surveys are conducted in the coming years, we expect our model to continue to provide good estimates given that it has proven to be flexible in responding to any shocks that may manifest as evidenced by examples in Figure 8. That said, this model specification would not immediately capture a trend break in school completion due to COVID-19 given that the drift parameter is shared by all years. Similarly, even though there have yet to be any signs that the adoption of the SDG agenda did actually induce a major trend break, if one was to occur, it would not manifest as an explicit change in long-term drift and instead would be found as a realignment though a shock in the residuals.

Two key challenges are identified. At a fundamental level, absent unbiassed sources of estimates for at least some countries, we can only estimate the relative bias of different surveys. The fact that all available surveys may be undersampling educationally distinctive population groups cannot at present be accounted for without incorporating strong a priori assumptions about this general bias. At this time, however, there is inadequate information to make such an assumption about absolute bias in a principled manner. In addition, for the 16 countries with only a single survey, the relative bias specification cannot attribute any survey bias uncertainty given the lack of consistent survey wave or type level patterns in the bias structure.

Secondly, while the principal rationale for the completion rate indicator was that individuals in the reference age brackets could be assumed to have completed the school level in question, our results show that cases of very late completion, even 5 years or more above the theoretical graduation age, are common. At the same time, however, both the amount and age pattern of late completion differs greatly across countries and levels of schooling. This represents a key challenge to the model specification, where all these patterns must be captured parsimoniously. There is a trade-off between fitting late completion and the ability to identify recent decline in ultimate completion. The current specification may err on the side of identifying declines, as several cases can be identified where, given contextual background knowledge, projected declines are recognised as spurious consequences of atypically late completion patterns.

At some level, the finding of such widespread and considerable late completion is a serious concern, because the rationale behind defining the completion rate indicator with respect to an age group several years above the nominal graduation age was precisely to minimise this effect. The true ultimate completion rate in a given cohort can be observed at some significantly higher age such as 30 years such that further school completion can safely be assumed to be statistically negligible. The age bracket 3–5 years above the nominal age for the final grade was assumed to be a reasonably good approximation. Our results clearly show that the completion rate indicator thus defined *cannot* be interpreted as a proxy for ultimate cohort completion. Instead, it should be recognised as measuring what might be termed 'reasonably timely' completion.

ACKNOWLEDGEMENTS

We would like to thank the members of the Technical Cooperation Group on the Indicators for SDG 4 for their input. Particular thanks go to Silvia Montoya, the director of the UNESCO Institute for Statistics, who embraced the concept and successfully made the case for the adoption of Indicator 4.1.2 to the Inter-agency and Expert Group on SDG Indicators. We also thank Monica Alexander for her comments and feedback throughout this project. This paper is based on data from the DHS, MICS, Integrated Public Use Microdata Series (IPUMS), Encuesta Permanente de Hogares (EPH), Integrated Living Conditions Survey (ILCS), Encuestas de Hogares (EH), Pesquisa Nacional por Amostra de Domicílios (PNAD), Encuesta de Caracterización Socioeconómica Nacional (CASEN), the China Family Panel Studies (CFPS), Encuesta Nacional de Calidad de Vida (ECV), Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), Encuesta de Hogares de Propositos Multiples (EHPM), Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), Household and Income Expenditure Survey (HIES), Encuesta Permanente de Hogares Continua (EPHC), Encuesta Nacional de Hogares (ENAHO), the Russia Longitudinal Monitoring Survey - Higher School of Economics (RLMS-HSE), Household Budget Survey (HBS), and Encuesta Continua de Hogares (ECH) from the years 1999–2019. We would also like to acknowledge the Cross-national Data Center in Luxembourg for its permission to use the Luxembourg Income Study (LIS) database and Eurostat for its permission to use the European Union Survey on Income and Living Conditions (EU-SILC). The responsibility for all conclusions drawn from the data lies entirely with the authors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from DHS, MICS, EU-SILC, LIS, IPUMS, EPH (Argentina), ILCS (Armenia), EH (Bolivia), PNAD (Brazil), CASEN (Chile), CFPS (China), ECV (Colombia), ENEMDU (Ecuador), EHPM (El Salvador), ENIGH (Mexico), HIES (Papua New Guinea), EPHC (Paraguay), ENAHO (Peru), RLMS-HSE (Russia), HBS (Tanzania), and ECH (Uruguay). Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://dhsprogram.com/, https://mics.unicef.org/, https://ec.europa.eu/eurostat/web/main/data/database, http://www.lisdatacenter.org, https://doi.org/10.18128/D020.V7.2, https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos, https://armstat.am/en/, https://www.ine.gob.bo/index.php/estadisticas-sociales/vivienda-y-servicios-basicos/encuestas-de-hogares-vivienda/, https://www.ibge.gov.br/en/statistics/social/labor/20620-summary-of-indicators-pnad2.html?=-&t=acesso-ao-produto,; http://observatorio.ministeriodesarrollosocial.gob.cl/index.php, https://opendata.pku.edu.cn/dataverse/CFPS?language=en, https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/calidad-de-vida-ecv, https://www.ecuadorencifras.gob.ec/enemdu-trimestral/, http://www.digestyc.gob.sv/index.php/temas/des/ehpm.html, https://www.inegi.

org.mx/programas/enigh/nc/2018/, https://www.nso.gov.pg/census-surveys/household-and-income-expenditure-survey/, https://www.ine.gov.py/, http://iinei.inei.gob.pe/microdatos/, https://rlms-hse.cpc.unc.edu/, https://www.nbs.go.tz/index.php/en/census-surveys/poverty-indicators-statistics/household-budget-survey-hbs, https://www.ine.gub.uy/encuesta-continua-de-hogares1 with the permission of DHS, MICS, EU-SILC, LIS, IPUMS, EPH, ILCS, EH, PNAD, CASEN, CFPS, ECV, ENEMDU, EHPM, ENIGH, HIES, EPHC, ENAHO, RLMS-HSE, HBS, and ECH, respectively.

ORCID

Ameer Dharamshi https://orcid.org/0000-0002-5505-4765

REFERENCES

Alexander, M. (2020) *Distortr: temporal smoothing methods for demographic time series*. Available at: https://github.com/MJAlexander/distortr.

Alexander, M. & Alkema, L. (2018) Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, 38, 335–372. Available from. https://doi.org/10.4054/DemRes.2018.38.15

Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A.B., Gemmill, A. et al. (2016) Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *The Lancet*, 387, 462–474. Available from. https://doi.org/10.1016/S0140-6736(15)00838-7

Alkema, L. & New, J.R. (2014) Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, 8, 2122–2149. Available from. https://doi.org/10.1214/14-aoas768

ArmStat. (2018) *Integrated living conditions survey [Datasets]*. National Statistical Service of the Republic of Armenia. Available at: https://armstat.am/en/.

Barro, R.J. & Lee, J.-W. (1993) International comparisons of educational attainment. *Journal of Monetary Economics*, 32, 363–394. Available from. https://doi.org/10.1016/0304-3932(93)90023-9

Bengtsson, H. (2020a) *A unifying framework for parallel and distributed processing in R using futures*. Available at: https://arxiv.org/abs/2008.00553.

Bengtsson, H. (2020b) *Future: unified parallel and distributed processing in R for everyone*. Available at: https://github.com/HenrikBengtsson/future.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software, Articles*, 76, 1–32. Available from. https://doi.org/10.18637/jss.v076.i01

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2009) Handling sparsity via the horseshoe. In: van Dyk, D. & Welling, M. (Eds.) *Proceedings of machine learning research*. Hilton Clearwater Beach Resort, Clearwater Beach, PMLR, Available at:. http://proceedings.mlr.press/v5/carvalho09a.html

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. & Liu, J. (2013) A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78, 685–709. Available from. https://doi.org/10.1007/s11336-013-9328-2

DANE. (2019) *Encuesta nacional de calidad de vida [Datasets]*. Departamento Administrativo Nacional de Estadística. Available at: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/calidad-de-vida-ecv.

DIGESTYC. (2019) *Encuesta de hogares de propositos multiples [Datasets]*. Dirección General de Estadística y Censosx. Available at: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/calidad-de-vida-ecv.

EFA Global Monitoring Report. (2015) *How long will it take to achieve universal primary and secondary education?* Technical background note for the framework for action on the post-2015 education agenda. UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000233028.

Eurostat. (2005–2017) *EU statistics on income and living conditions (various) [Datasets]*. Eurostat. Available at: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. (2019) Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 389–402. Available from. https://doi.org/10.1111/rssa.12378

Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472. Available at:. http://www.jstor.org/stable/2246093

Hoffman, M.D. & Gelman, A. (2014) The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623 Available at: http://jmlr.org/papers/v15/hoffman14a.html

HSE. (2000–2019) *Russia longitudinal monitoring survey - Higher school of economics [Datasets]*. National Research University Higher School of Economics; ZAO 'Demoscope' together with Carolina Population Center, University of North Carolina at Chapel Hill; the Institute of Sociology RAS. Available at: https://rlms-hse.cpc.unc.edu/.

IBGE. (2007–2015) *Pesquisa Nacional por Amostra de Domicílios [Datasets]*. Instituto Brasileiro de Geografia e Estatística. Available at: https://www.ibge.gov.br/en/statistics/social/labor/20620-summary-of-indicators-pnad2.html?=&t=acesso-ao-produto.

ICF. (2000–2018) *Demographic and health surveys (various) [Datasets]*. Rockville, MD: ICF; Funded by USAID. Available at: https://dhsprogram.com/

INDEC. (2004–2012) *Encuesta Permanente de Hogares [Datasets]*. Instituto Nacional de Estadística y Censos de la República Argentina. Available at: https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos.

Independent Expert Advisory Group on the Data Revolution for Sustainable Development. (2014) *A world that counts: mobilising the data revolution for sustainable development*. New York, NY: UN Data Revolution. Available at. https://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf

INE. (2019a) *Encuesta Continua de Hogares [Datasets]*. Instituto Nacional de Estadística. Available at: https://www.ine.gub.uy/encuesta-continua-de-hogares1.

INE. (2019b) *Encuesta Permanente de Hogares Continua [Datasets]*. Instituto Nacional de Estadística. Available at: https://www.ine.gov.py/.

INE. (2019c) *Encuestas de Hogares [Datasets]. Instituto Nacional de Estadística*. Available at: https://www.ine.gob.bo/index.php/estadisticas-sociales/vivienda-y-servicios-basicos/encuestas-de-hogares-vivienda/.

INEC. (2018) *Encuesta Nacional de Empleo, Desempleo y Subempleo [Datasets]*. Instituto Nacional de Estadística y Censos. Available at: https://www.ecuadorencifras.gob.ec/enemdu-trimestral/.

INEGI. (2018) *Encuesta Nacional de Ingresos y Gastos de los Hogares [Datasets]*. Instituto Nacional de Estadística, Geografía e Informática. Available at: https://www.inegi.org.mx/programas/enigh/nc/2018/.

INEI. (2019) *Encuesta Nacional de Hogares [Datasets]*. El Instituto nacional de Estadística e Informática. Available at: http://iinei.inei.gob.pe/microdatos/.

Kozyreva, P., Kosolapov, M. & Popkin, B.M. (2016) Data resource profile: the Russia longitudinal monitoring survey—Higher school of economics (RLMS-HSE) Phase II: monitoring the economic and health situation in Russia, 1994–2013. *International Journal of Epidemiology*, 45, 395–401. Available from. https://doi.org/10.1093/ije/dyv357

Landau, W.M. (2018) The drake R package: A pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, 3, 550. Available from. https://doi.org/10.21105/joss.00550

Landau, W.M. (2020) *Drake: a pipeline toolkit for reproducible computation at scale*. Available at: https://CRAN.R-project.org/package=drake.

Luxembourg Income Study (LIS) Database. (2005–2017) *Luxembourg income study (LIS) database (multiple countries; 2001-2018) [Datasets]*. Luxembourg: LIS. Available at: http://www.lisdatacenter.org

Malala Fund. (2014) *Malala fund is working for a world where every girl can learn and lead*. Washington, D.C.: Malala Fund. Available at: https://malala.org/about?sc=header

Ministerio de Desarrollo Social y Familia. (2000–2015) *Encuesta de Caracterización Socioeconómica Nacional [Datasets]. Ministerio de Desarrollo Social y Familia*. Available at: http://observatorio.ministeriodesarrollosocial.gob.cl/index.php.

Minnesota Population Center. (2020) *Integrated public use microdata series, international: version 7.3 [Datasets]*. Minneapolis, MN: IPUMS. Available from. https://doi.org/10.18128/D020.V7.2

Neal, R.M. (2011) MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. (Eds.) *Handbook of Markov Chain Monte Carlo*. Boca Raton: Chapman & Hall/CRC Press. Available from. https://doi.org/10.1201/b10905

NSO. (2009) *Household and Income Expenditure Survey [Datasets]*. National Statistical Office of Papua New Guinea. Available at: https://www.nso.gov.pg/census-surveys/household-and-income-expenditure-survey/.

Peking University Open Research Data. (2010–2014) *China family panel studies [Datasets]*. Peking University Open Research Data. Available at: https://opendata.pku.edu.cn/dataverse/CFPS?language=en.

Piironen, J. & Vehtari, A. (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11, 5018–5051. Available from: https://doi.org/10.1214/17-ejs1337si

Schubert, M. (2019) clustermq enables efficient parallelization of genomic analyses. *Bioinformatics*, 35, 4493–4495. Available from: https://doi.org/10.1093/bioinformatics/btz284

Schubert, M. (2020) *Clustermq: evaluate function calls on HPC schedulers (Lsf, Sge, Slurm, Pbs/Torque)*. Available at: https://mschubert.github.io/clustermq/.

Stan Development Team. (2020a) *Brief guide to Stan's warnings*. Available at: https://mc-stan.org/misc/warnings.html.

Stan Development Team. (2020b) *Prior choice recommendations*. Available at: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Stan Development Team. (2020c) *Rstan: R interface to Stan*. Available at: https://CRAN.R-project.org/package=rstan.

The Demographic and Health Surveys Program. (2012) *Demographic and health survey sampling and household listing manual*. ICF International. Available at: https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

TNBS. (2017) *Household budget survey [Datasets]. Tanzania National Bureau of Statistics*. Available at: https://www.nbs.go.tz/index.php/en/census-surveys/poverty-indicators-statistics/household-budget-survey-hbs.

UN Statistical Division. (2021a) *Indicator 4.1.1: proportion of children and young people (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex*. UN Statistical Division. Available at: https://unstats.un.org/sdgs/metadata/files/Metadata-04-01-01.pdf.

UN Statistical Division. (2021b) *Indicator 4.1.2: completion rate (primary education, lower secondary education, upper secondary education)*. UN Statistical Division. Available at: https://unstats.un.org/sdgs/metadata/files/Metadata-04-01-02.pdf.

UNESCO. (2012) *Youth and skills: putting education to work*. UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000218003.

UNESCO. (2016) *Global education monitoring report 2016 box 14.2*. UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000245752.

UNESCO Institute for Statistics, Global Education Monitoring Report Team. (2019) *Meeting commitments: are countries on track to achieve sdg 4?* UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000369009.

UNICEF. (1999–2019) *Multiple indicator cluster survey (various) [Datasets]*. UNICEF. Available at: https://mics.unicef.org/.

United Nations. (2015) *Transforming our world: the 2030 agenda for sustainable development*. United Nations. Available at: https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981.

United Nations. (2022) SDG indicators database. Available at: https://unstats.un.org/sdgs/dataportal.

Vehtari, A., Gabry, J., Yao, Y. & Gelman, A. (2020) *Loo: efficient leave-one-out cross-validation and WAIC for Bayesian models*. Available at: https://CRAN.R-project.org/package=loo.

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. Available from: https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.C. (2020) Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16, 667–718. Available from: https://doi.org/10.1214/20-ba1221

World Bank, International Monetary Fund. (2011) *Global monitoring report 2011: improving the odds of achieving the mdgs*. World Bank. Available at: https://openknowledge.worldbank.org/handle/10986/2293.

Xie, Y. & Lu, P. (2015) The sampling design of the China family panel studies (CFPS). *Chinese Journal of Sociology*, 1, 471–484. Available from: https://doi.org/10.1177/2057150X15614535

Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. (2017) Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, 13, 917–1007. Available from: https://doi.org/10.1214/17-BA1091

How to cite this article: Dharamshi, A., Barakat, B., Alkema, L. & Antoninis, M. (2022) A Bayesian model for estimating Sustainable Development Goal indicator 4.1.2: School completion rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5), 1822–1864. Available from: https://doi.org/10.1111/rssc.12595

APPENDIX A. MODEL DETAILS

Putting all errors and adjustments together, we model the empirical observations K_i of probit completion relating to age $a[i]$, country $c[i]$, year $y[i]$ and originating from survey $s[i]$ as resulting from: the 'true' probit completion $\kappa_{c[i],y[i]}$, survey bias $\beta_{s[i]}$, a distortion due to age-misreporting $\tau_{c[i]}$ and the late (relative to a_5) completion term $\phi_{a[i],c[i]}$, and a total error variance consisting of sampling and non-sampling error with variances ν_i^2 and $\omega^2_{a[i],s[i]}$, respectively:

$$K_i|\kappa_{c[i],y[i]}, \beta_{s[i]}, \tau_{c[i]}, \phi_{a[i],c[i]}, \nu_i, \omega_{a[i],s[i]} \sim \mathcal{N}(\kappa_{c[i],y[i]} + \beta_{s[i]} - \tau_{c[i]} \cdot \mathbb{1}_{5|a[i]} + \phi_{a[i],c[i]}, \quad \nu_i^2 + \omega^2_{a[i],s[i]}), \tag{A1}$$

The model can be divided into two stages. First is the process model for the underlying 'true' completion rates, $\kappa_{c[i],y[i]}$. The second details how the underlying completion rates relate to the observed data. This is divided into adjustments for late completion and those for various data considerations.

Our estimation is conducted within a Bayesian framework. For the most part, we assign vaguely informative priors. The following is a discussion of the rationale behind our choices.

A.1 Process model

We assume the underlying 'true' values follow an ARIMA(1,1,0) process with drift:

$$\Delta\kappa_{c,y} = \kappa_{c,y} - \kappa_{c,y-1} = \gamma_c + \rho_c \Delta\kappa_{c,y-1} + \epsilon_{c,y}, \tag{A2}$$

$$\epsilon_{c,y}|\sigma_\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{A3}$$

with the following priors:

$$\sigma_\epsilon \sim \text{Gamma}(2, 0.1), \tag{A4}$$

$$\rho_c|\tau, \lambda_c \sim \mathcal{N}^+(0, \tau^2\lambda_c^2), \tag{A5}$$

$$\tau|\tau_0 \sim \mathcal{N}^+(0, \tau_0^2), \tag{A6}$$

$$\lambda_c \sim t_4^+(0, 1) \tag{A7}$$

$$\gamma_c|\alpha_\gamma, \beta_\gamma \sim \text{Log-normal}(\alpha_\gamma, \beta_\gamma), \tag{A8}$$

$$\alpha_\gamma \sim \mathcal{N}(0,1), \tag{A9}$$

$$\beta_\gamma \sim \mathcal{N}^+(0,1). \tag{A10}$$

With respect to the country-specific drift γ_c, we specify a log-normal distribution with a hyperprior setup. The choice of a distribution constrained to be non-negative reflects the expectation of a positive trend in the long run, in line with the objectives of SDG 4. Further, for a country currently far from zero, a negative long-term drift implying eventual convergence to zero completion is not plausible.

Any short- to medium-term negative trends are then modeled with $\epsilon_{c,y}$ and the autoregressive coefficient ρ_c. Given that significant departures from the long-term trend are not expected in all countries, we use a horseshoe prior on ρ_c to shrink the ρ_c terms corresponding to the countries without these irregular patterns to zero. The horseshoe prior setup here uses a t-distribution for λ_c instead of the standard Cauchy distribution to simplify for computation purposes. The lighter tails of the country-specific shrinkage distribution is compensated by a higher variance of the distribution of τ, the global shrinkage distribution. The hyperparameter for the prior variance of τ, τ_0, is selected following the discussion in Piironen and Vehtari (2017). There, the shrinkage factor is defined as $\kappa_j = (1 + n\sigma^{-2}\tau^2\lambda_j^2)^{-1}$ and the effective number of nonzero coefficients is $m_{eff} = \sum_{j=1}^{D}(1-\kappa_j)$ where n is the sample size, σ is the error and D is the dimension. If $\lambda_j \sim \text{Cauchy}^+(0,1)$, setting the expected number of non-zero coefficients to a prior guess p_0 yields the value $\tau_0 = \frac{p_0}{D-p_0}\frac{\sigma}{\sqrt{n}}$ for the τ hyper-parameter. Given that we use $\lambda_c \sim t_4^+(0,1)$ instead of $\lambda_c \sim \text{Cauchy}^+(0,1)$ we simulated values to κ to select the value $\tau_0 = 0.01$ that corresponds to approximately 30% as the prior guess for the number of non-zero coefficients.

We allow the scaling of year-over-year residuals to be determined by the data though the prior on said scaling σ_ϵ is specifically boundary avoiding to prevent a collapsing scenario. For perspective on what the effects of a given magnitude in the transformed space imply on the outcome scale of percent completion, note that for $C = 0.5$, that is, 50% completion in year y, a 1 percentage point change corresponds approximately to a change in κ of ± 0.025. Practically speaking, while the Gamma$(2, 0.1)$ prior is well overdispersed, the resulting scaling terms are closer in scale to 0.02, an entirely plausible value for the jitter term while still allowing for reasonably large shocks.

However, we do acknowledge that in the extreme tails of the probit curve, small amounts of noise in the real space translate to seismic shocks in the probit space. For example, values of 4 and 5 in the probit space correspond to observations of 99.997% completion and 99.99997% completion respectively. Ultimately, these values are both indicating universal completion in the real space. However, when modelling in the probit, space, what is really just noise appears to be dramatic shifts in completion. To avoid attempting to model the extreme noise, we impose a cap on extreme values such that if a country observes its maximum transformed value above 2.5 or minimum transformed value below -2.5 (99.4% and 0.06% completion, respectively), all of its observations are uniformly shifted inwards such that the maximum is now 2.5 or minimum is -2.5, respectively. This reduces the risk of noise having undue influence on model parameters. In post-processing, the extracted true values are shifted back to restore the original levels.

The initial value of κ is assigned an uninformative prior, $\kappa_{c,1980} \sim N(0, 10)$ to allow the data to determine the intercept.

A.2 Late completion

The age profile is captured in $\phi_{a,c}$:

$$\phi_{a,c} = \begin{cases} (a - a_5) \cdot \lambda_{1c} \cdot \mathbb{1}_l & \text{if } a \in \{a_3, a_4\} \\ \min(3, a - a_5) \cdot \lambda_{2c} \cdot \mathbb{1}_{vl} & \text{if } a \geq a_5 \end{cases}. \tag{A11}$$

$$\lambda_{1c} | \sigma_{\lambda_1} \sim \mathcal{N}^+(0, \sigma^2_{\lambda_1}). \tag{A12}$$

$$\lambda_{2c} | \sigma_{\lambda_2} \sim \mathcal{N}^+(0, \sigma^2_{\lambda_2}). \tag{A13}$$

$$\sigma_{\lambda_1} \sim \mathcal{N}^+(0, 1). \tag{A14}$$

$$\sigma_{\lambda_2} \sim \mathcal{N}^+(0, 1). \tag{A15}$$

A visual equivalent to the $\phi_{a,c}$ specification is provided in Figure 7. The indicators in $\phi_{a,c}$ reflect the reality that late completion is only estimated if there is an indication of its presence being more than simply noise. Domain knowledge suggests that in highly developed education systems with close-to-universal completion, significant amounts of completion several years above the standard age are extremely unlikely and should be interpreted as data problems. That is, the grace period offered by the completion rate indicator is considered to successfully capture all reasonably short delays in completion. Here we define close-to-universal completion as a median observed completion rate above 0.95. In the case of the long-delayed completion parameter, λ_{2c}, it is only estimated for those countries with median observed values below 0.95. In the case of the medium-delay completion parameter, λ_{1c}, it is estimated for those countries with median observed values below 0.95 *or* $[a_3, a_5]$ values consistently below $[a_5, a_7]$. In other words, we allow for structural medium-delayed completion to be a possibility even in countries with close-to-universal completion if the data suggests that is the case. Late completion $\phi_{a,c}$ is a 'real' effect in the sense that the true completion at ages other than a_5 really is different and this is not a measurement artefact.

A.3 Data considerations

Survey Bias

There are a number of inherent challenges regarding survey bias using household surveys. If *all* surveys overestimate school completion, for example because they exclude street children, this shared bias cannot be identified without additional assumptions and/or data. Accordingly, if one survey is actually unbiassed, and another biassed, but we cannot identify which is unbiassed, the model estimate will attenuate the latter bias, but will also 'correct' the relative 'bias' of the former. In other applications of similar models, this is partly remedied either by exploiting prior information regarding the absolute bias of specific surveys (gained from an intensive re-count in a sub-sample, for instance), or by comparison with a 'gold standard' data source that is assumed to suffer a low bias.

The lack of a gold standard in the education context precludes the estimation of the *absolute* bias in survey-based estimates. Nevertheless, modelling the bias of available surveys *relative to each other* allows for an unbiassed estimation of what would be estimated if surveys of all type were available for every year, even when only a subset or only a single survey actually is. In other words, if series A were consistently lower than series B, then for years in which only observations from series A are available, we may still conclude that this is likely to be an

underestimate, and that the model estimate should be higher. We thus settle on a relative survey bias structure.

Define β_c to be the vector of S survey biases β_s related to country c when $S > 1$. Note that when $S = 1$, there can be no relative survey bias and thus the estimate for the single survey's bias is zero. Next, let β_c^* be a vector of length $S - 1$ attributed to country c. Then, we construct the survey bias estimates as follows:

$$\beta_c = A_S \cdot \beta_c^*. \tag{A16}$$

$$A_S = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\frac{1}{S-1} & 1-\frac{1}{S-1} & 0 & \dots & 0 \\ -\frac{1}{S-1} & -\frac{1}{S-1} & 1-\frac{2}{S-1} & & 0 \\ \vdots & \vdots & & \ddots & \\ -\frac{1}{S-1} & -\frac{1}{S-1} & \dots & -\frac{1}{S-1} & 1-\frac{S-2}{S-1} \\ -\frac{1}{S-1} & -\frac{1}{S-1} & -\frac{1}{S-1} & \dots & -\frac{1}{S-1} \end{bmatrix}. \tag{A17}$$

$$\beta_c^* | \sigma_{\text{bias}} \sim \text{Cauchy}(0, \sigma_{\text{bias}}). \tag{A18}$$

$$\sigma_{\text{bias}} \sim \mathcal{N}^+(0, 0.25^2). \tag{A19}$$

The relative structure of the survey bias induces a sum-to-zero behaviour in the survey bias terms for each country. Equivalently, if one were to directly estimate S survey bias terms for a country with S surveys, one term would be redundant. This redundancy produces ridge-like geometry which impedes sampling in the Bayesian framework. Instead, parameterising in terms of β_c^*, a vector of $S - 1$ values, and transforming into the S survey bias terms with the $S \times (S - 1)$ A_S matrix reduces the degrees of freedom by one as demanded by the sum-to-zero constraint.

The A_S matrix used here is designed to serve two distinct purposes. First define a_{ij} to be the elements of A_S. Notice that for all columns j, $\sum_{i=1}^{S} a_{ij} = 0$, thus enforcing the sum-to-zero constraint in β_c. Second, notice that for all rows i, $\sum_{j=1}^{S-1} |a_{ij}| = 1$. This property propagates the Cauchy$(0, \sigma_{\text{bias}})$ prior on β_c^* to an implied prior on β_c using the properties of the Cauchy distribution. The Cauchy distribution has been selected to capture the possibility of extreme outlier surveys, a possibility that is observed in the data. An example of such a survey is presented for Armenia in Figure 9. The model recognises that the highly biassed or erroneous input could be offering relevant information but does not dramatically deviate from the rest of the surveys.

Age-misreporting distortion

The error in observed completion rates for ages divisible by 5 due to age-misreporting, τ_c, has the following prior:

$$\tau_c | \lambda_\tau \sim Exp(\lambda_\tau). \tag{A20}$$

$$\lambda_\tau \sim \mathcal{N}^+(0, 50^2). \tag{A21}$$

Variance

The model accounts for the total error variances resulting from the combination of sampling and non-sampling errors, such that observations of completion rates with larger total error variance carry less weight.

The sampling variance v_i^2 of a given specific observation is estimated by clustered Jackknife prior, as input into the model. Specifically, the sampling variance of any given observed completion rate $C_{a,y,c,s}$ in year y at age a in country y from survey s is estimated as (omitting indices for clarity):

$$\widehat{\text{Var}}(C) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (C_i - C)^2, \tag{A22}$$

where

$$C_i = nC - (n-1)C_{(i)}. \tag{A23}$$

Here, C is calculated on the full sample, $C_{(i)}$ is calculated on the sample with the ith cluster excluded, and n is the total number of clusters. Given the presence of 100% completion observations in the data, we calculate C with a minor jitter sourced from a Beta(0.5, 0.5) prior that serves the dual purposes of moving observations slightly off 1 for the probit transformation as well as creating microscopic (importantly non-zero) variance when using the jackknife. For IPUMS data, in light of the fact that standard errors of the census samples are in any case much smaller than of the surveys, for simplicity the same approach was applied, with 1000 random 'clusters', instead of customising the process to the specific stratification of each sample.

After computing the sampling variance in the observed space, it is transformed to the probit space using the delta method as:

$$v_i^2 = \frac{\widehat{\text{Var}}(C)}{\left(f(\Phi^{-1}(C))\right)^2}, \tag{A24}$$

where f and Φ^{-1} are the density and inverse CDF of the standard normal distribution, respectively.

The extent to which observations *do* differ with respect to sampling variability across countries, but crucially also across ages, time, and surveys within countries, is shown in Figure A1 for a country with a typical (Mali) and a wide (Belize) spread of estimated SEs. The conclusion is that not all data points call for an equally close fit by the model. Even for Belize, if the fitted trend missed an observation by 4 percentage points, say, this would stretch credulity for some observations, but could perfectly plausibly be attributed to sampling error for others.

Non-sampling variance $\omega_{a,s}^2$ is composed of a base variance ω_s^2 and an inflation factor capturing increased uncertainty due to reconstruction:

$$\omega_{a,s}^2 = (1 + 0.05 \cdot \max(0, a - a_5)) \cdot \omega_s^2. \tag{A25}$$

$$\omega_s \sim \text{Gamma}(2, 4). \tag{A26}$$

The gamma prior is selected for its boundary avoiding properties. Unlike with total variance, a zero value for non-sampling variance could be consistent with the likelihood given that sampling

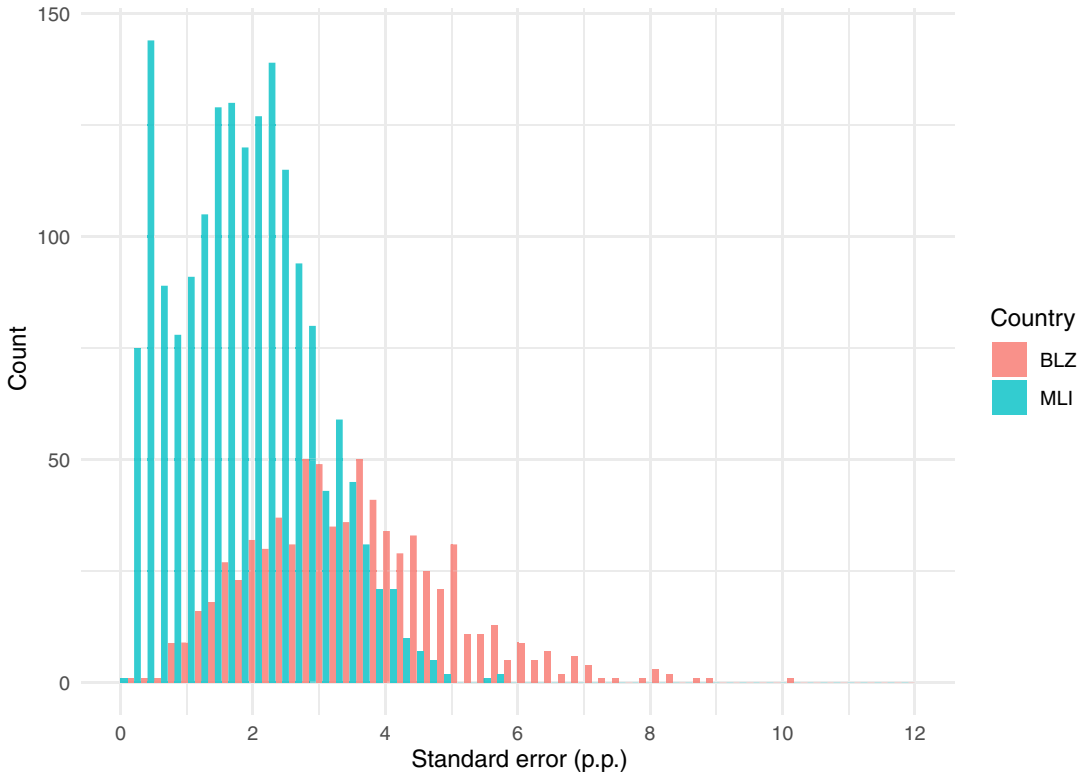

FIGURE A1 Distribution of estimated sampling standard errors for age-specific completion rates, example countries of Belize and Mali [Colour figure can be viewed at wileyonlinelibrary.com]

variance is guaranteed to be positive. The gamma distribution reflects the understanding that non-sampling variance is certainly present.

A.4 Population weights

The customary completion rate indicator is the average completion rate over empirically observed individuals in the 3-year age interval $[a_3, a_5]$, in other words the implicitly population-weighted average. Based on empirically observed completion, it is:

$$\mathrm{CR}_{c,y} = C_{[a_3,a_5],c,y} = \sum_{i=3}^{5} \frac{p_i}{p_{[3,5]}} C_{a_i,c,y}. \tag{A27}$$

Here, p_i is the size of the observed population aged a_i, i years above the last grade of a given level of schooling, and $p_{[3,5]}$ is the overall population in the age interval $[a_3, a_5]$.

Consider a uniform random sample of size 20,000 and a true single year cohort share of 2%, the binomial standard error of the sampled single year cohort share would be $\frac{\sqrt{20{,}000 \cdot 0.02 \cdot 0.98}}{20{,}000} \approx 0.001$, or 5% in relation to the true value of 0.02. By comparison, only the most extreme cohort-on-cohort growth rates reach 3%, suggesting that random variation in the age distribution will significantly exceed true differences in birth cohort size in all but the largest surveys and extreme fertility settings.

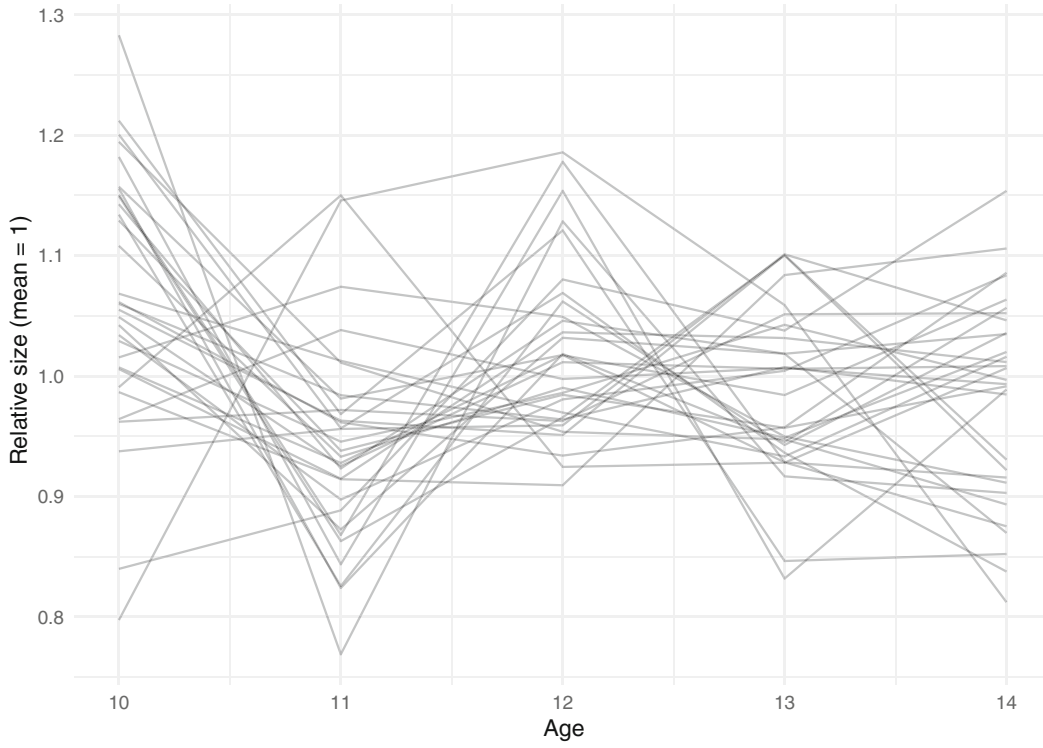

FIGURE A2 Relative (weighted) size of single-year age groups 10 to 14 in MICS5 surveys

This can be confirmed empirically, as in Figure A2, which for illustration shows the relative size of individual age cohorts in the age interval 10–14 in surveys from the MICS5 series. These profiles are clearly dominated by random fluctuations rather than smooth trends in cohort size driven by population growth. Year on year fluctuations of up to 25% are unlikely to represent differences in true cohort size. Moreover, differences in the sizes of single year *retrospective* cohorts will further be distorted by random variation in mortality and migration.

The unweighted average is, therefore, arguably preferable in general, and is certainly more suitable for modelling and projection. This age-standardised completion rate indicator $\mathrm{CR}^*_{c,y}$ is:

$$\mathrm{CR}^*_{c,y} = \frac{1}{3}\sum_{a=a_3}^{a_5} C_{a,c,y}. \tag{A28}$$

The equivalent definition for the *true* completion rates is:

$$\widehat{\mathrm{CR}^*}_{c,y} = \frac{1}{3}\sum_{a=a_3}^{a_5} \Gamma_{c,y} = \frac{1}{3}\sum_{a=a_3}^{a_5} \Phi(\kappa_{c,y} + \phi_{a,c}). \tag{A29}$$

A.5 Validation

The complete validation results for all education levels and populations are presented here. The results from the 'leave one survey out' exercise are in Table A1, Table A2 contains the results for the random test set exercise, and Table A3 contains the corresponding coverage values.

TABLE A1 Full adjusted Bayesian completion rates (ABCR) 'leave one survey out' validation results

Error (×100)	ABCR	Simple	ABCR (no late)	ABCR (no bias)	ABCR (no distortion)
Primary/female					
MSE	0.28	0.43	0.62	0.24	0.30
MAE	3.08	3.77	4.31	2.88	3.16
Primary/male					
MSE	0.32	0.54	1.37	0.29	0.33
MAE	3.39	4.37	6.32	3.27	3.44
Primary/total					
MSE	0.26	0.44	0.87	0.23	0.28
MAE	3.00	3.91	5.01	2.83	3.08
Lower secondary/female					
MSE	0.48	0.64	0.70	0.43	0.48
MAE	4.26	4.96	5.15	4.05	4.28
Lower secondary/male					
MSE	0.71	0.89	1.26	0.64	0.71
MAE	5.34	5.90	7.14	5.07	5.36
Lower secondary/total					
MSE	0.53	0.89	0.91	0.51	0.53
MAE	4.50	5.74	5.94	4.41	4.50
Upper secondary/female					
MSE	0.83	0.95	0.98	0.87	0.84
MAE	6.09	6.57	6.75	6.07	6.12
Upper secondary/male					
MSE	1.09	1.22	1.40	1.17	1.10
MAE	7.40	7.91	8.60	7.52	7.44
Upper secondary/total					
MSE	0.88	1.01	1.12	0.96	0.89
MAE	6.32	6.89	7.43	6.45	6.37

TABLE A2 Full adjusted Bayesian completion rates (ABCR) random test set validation results

Error (×100)	ABCR	Simple	ABCR (no late)	ABCR (no bias)	ABCR (no distortion)
Primary/female					
MSE	0.09	0.30	0.16	0.16	0.10
MAE	1.73	3.21	2.30	2.23	1.89
Bias	−0.02	−0.11	0.04	−0.18	−0.08
Primary/male					
MSE	0.09	0.44	0.22	0.15	0.11
MAE	1.77	3.85	2.60	2.34	1.96
Bias	0.14	0.18	0.18	−0.08	0.13
Primary/total					
MSE	0.07	0.33	0.18	0.14	0.09
MAE	1.52	3.19	2.25	2.11	1.71
Bias	0.11	0.25	0.18	−0.03	0.13
Lower secondary/female					
MSE	0.10	0.35	0.13	0.28	0.12
MAE	2.06	3.81	2.30	3.10	2.22
Bias	0.10	0.30	0.09	−0.09	0.09
Lower secondary/male					
MSE	0.12	0.52	0.21	0.36	0.15
MAE	2.35	4.77	2.97	3.50	2.56
Bias	0.13	0.29	0.11	−0.10	0.20
Lower secondary/total					
MSE	0.08	0.39	0.13	0.26	0.10
MAE	1.84	4.03	2.25	2.96	2.05
Bias	0.16	0.35	0.22	0.00	0.13
Upper secondary/female					
MSE	0.13	0.38	0.14	0.38	0.14
MAE	2.37	4.10	2.46	3.59	2.47
Bias	0.00	0.00	0.03	−0.27	−0.04
Upper secondary/male					
MSE	0.16	0.52	0.19	0.53	0.18
MAE	2.78	5.10	3.02	4.45	2.91
Bias	0.03	−0.05	0.06	−0.50	0.00
Upper secondary/total					
MSE	0.09	0.39	0.11	0.41	0.10
MAE	2.07	4.22	2.27	3.59	2.18
Bias	−0.03	−0.15	−0.04	−0.74	−0.07

TABLE A3 Full adjusted Bayesian completion rates (ABCR) random test set coverage results

Coverage level (%)	ABCR	Simple	ABCR (no late)	ABCR (no bias)	ABCR (no distortion)
Primary/female					
0.80	0.89	0.92	0.90	0.88	0.89
0.90	0.96	0.96	0.95	0.95	0.95
0.95	0.98	0.98	0.96	0.98	0.98
Primary/male					
0.80	0.89	0.92	0.89	0.88	0.88
0.90	0.94	0.97	0.94	0.95	0.94
0.95	0.97	0.98	0.97	0.97	0.97
Primary/total					
0.80	0.89	0.93	0.90	0.88	0.89
0.90	0.95	0.97	0.96	0.95	0.96
0.95	0.98	0.98	0.97	0.98	0.98
Lower secondary/female					
0.80	0.87	0.90	0.87	0.86	0.87
0.90	0.94	0.97	0.94	0.93	0.93
0.95	0.96	0.99	0.96	0.97	0.97
Lower secondary/male					
0.80	0.86	0.90	0.87	0.85	0.85
0.90	0.94	0.96	0.94	0.93	0.94
0.95	0.97	0.98	0.96	0.97	0.97
Lower secondary/total					
0.80	0.87	0.90	0.88	0.85	0.87
0.90	0.94	0.95	0.94	0.94	0.94
0.95	0.97	0.97	0.96	0.97	0.97
Upper secondary/female					
0.80	0.85	0.92	0.87	0.87	0.85
0.90	0.93	0.97	0.94	0.95	0.93
0.95	0.98	0.98	0.97	0.98	0.97
Upper secondary/male					
0.80	0.86	0.92	0.85	0.85	0.85
0.90	0.94	0.96	0.93	0.94	0.94
0.95	0.97	0.98	0.96	0.98	0.97
Upper secondary/total					
0.80	0.85	0.90	0.85	0.86	0.86
0.90	0.95	0.96	0.94	0.95	0.94
0.95	0.98	0.98	0.97	0.98	0.98

APPENDIX B. IMPLEMENTATION DETAILS

B.1 Data sources

The analysis is based on a consolidated collection of individual-level micro-data on school completion. The results presented here are based on 696 distinct surveys from 164 countries.

Specifically, sources include DHS (ICF, 2000–2018), Multiple Indicator Cluster Surveys (MICS) (UNICEF, 1999–2019), European Union Statistics on Income and Living Conditions (EU-SILC) (Eurostat, 2005–2017), Luxembourg Income Study (LIS) (Luxembourg Income Study (LIS) Database, 2005–2017), selected other non-standard household surveys that form the basis for the *World Inequality Database on Education* (WIDE), and international census samples from the Integrated Public Use Microdata Series (IPUMS) (Minnesota Population Center, 2020). For computational reasons, the IPUMS extracts were limited to 1 million observations each.

Returning to the category of non-standard surveys included in the present analysis, we examined the sample design and microdata of each of the surveys leveraged to ensure compatibility with the standard surveys. The specific sets of non-standard surveys are:

1. Argentina: Encuesta Permanente de Hogares (EPH) (INDEC, 2004–2012)
2. Armenia: Integrated Living Conditions Survey (ILCS) (ArmStat, 2018)
3. Bolivia: Encuestas de Hogares (EH) (INE, 2019c)
4. Brazil: Pesquisa Nacional por Amostra de Domicílios (PNAD) (IBGE, 2007–2015)
5. Chile: Encuesta de Caracterización Socioeconómica Nacional (CASEN) (Ministerio de Desarrollo Social y Familia, 2000–2015)
6. China: China Family Panel Studies (CFPS) (Peking University Open Research Data, 2010–2014; Xie & Lu, 2015)
7. Colombia: Encuesta Nacional de Calidad de Vida (ECV) (DANE, 2019)
8. Ecuador: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) (INEC, 2018)
9. El Salvador: Encuesta de Hogares de Propositos Multiples (EHPM) (DIGESTYC, 2019)
10. Mexico: Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) (INEGI, 2018)
11. Papua New Guinea: Household and Income Expenditure Survey (HIES) (NSO, 2009)
12. Paraguay: Encuesta Permanente de Hogares Continua (EPHC) (INE, 2019b)
13. Peru: Encuesta Nacional de Hogares (ENAHO) (INEI, 2019)
14. Russia: The Russia Longitudinal Monitoring Survey - Higher School of Economics (RLMS-HSE) (HSE, 2000–2019; Kozyreva et al., 2016)
15. Tanzania: Household Budget Survey (HBS) (TNBS, 2017)
16. Uruguay: Encuesta Continua de Hogares (ECH) (INE, 2019a)

It is worth noting that many of the countries listed have limited observations sourced through DHS, MICS and censuses. As such, the addition of non-standard surveys addresses the risk of limited insight on relative sample bias that results from limited survey data. That said, these additional surveys tend to have lower influence in the model due to higher sampling variance resulting from smaller sample sizes and in some cases, less robust sample designs.

B.2 Computation

The model was implemented and run in R version 4.0.2 (2020-06-22) calling on Stan version 2.21.0 on a x86_64-pc-linux-gnu (64-bit) platform. The present exercise draws inspiration from the `distortr` package (Alexander, 2020) by Monica Alexander that underpins the similarly-motivated models for infant and maternal mortality.

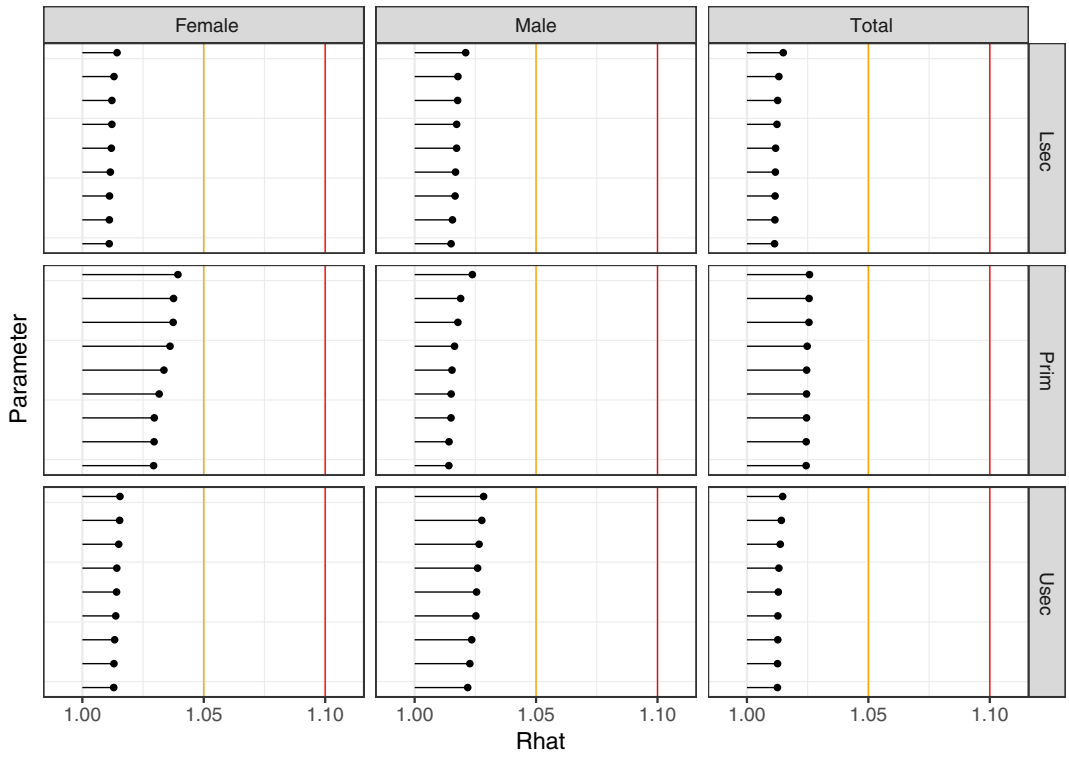

FIGURE B1 Highest potential scale reduction factors by level and sex [Colour figure can be viewed at wileyonlinelibrary.com]

Four chains were run for 3000 iterations each after having discarded 3000 iterations as burn-in, and thinned to 1000 iterations for the computations to reduce the memory footprint. The average runtime for a single chain across all level and sex combinations was 20.76 h. Chains were run in parallel using the `future` and `clustermq` packages (Bengtsson, 2020a, 2020b; Schubert, 2019, 2020), and the `drake` package (Landau, 2018, 2020) was used to ensure reproducibility and assist with version control.

B.3 Convergence

Convergence was assessed both through Gelman's criterion for *potential scale reduction factors* (PSRF) and visual inspection of parameter traceplots. We note that values below 1.1 for the PSRF are considered acceptable though the highest PSRF across model runs is 1.04. Figure B1 displays the PSRF values of the nine parameters with the highest (worst) PSRFs for the estimations for each level and total, female or male rates. The values are well below the acceptable thresholds. Sample traceplots for the 'worst' parameters corresponding to the model for primary education and both sexes are shown in Figure B2.

B.4 Validation

In addition to the out-of-sample validation exercises described in Section 4.1, we also considered in-sample performance and an approximate leave-one-out cross-validation (LOO-CV) exercise to provide additional insight on the quality of fit.

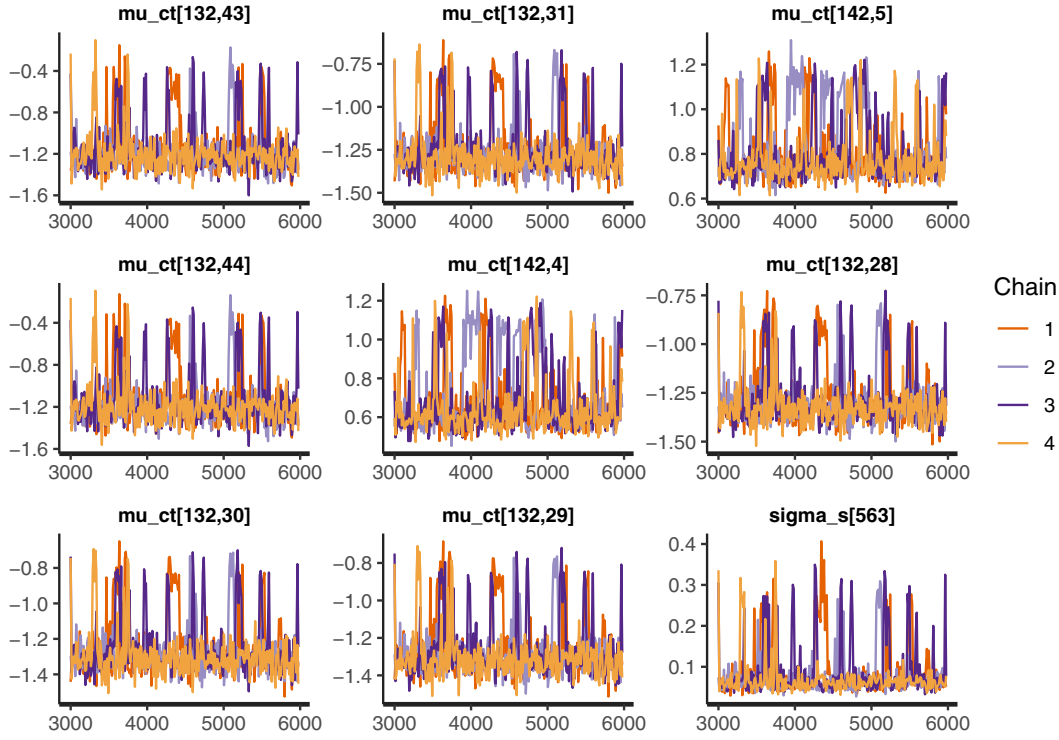

FIGURE B2 Example traceplots [Colour figure can be viewed at wileyonlinelibrary.com]

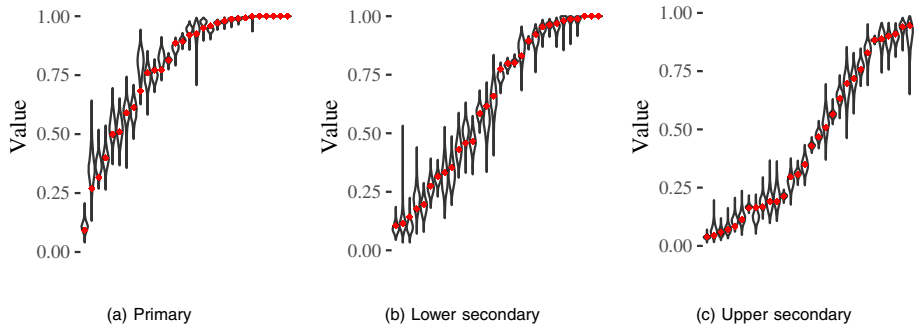

FIGURE B3 Posterior predictive distribution for a random sample of observation [Colour figure can be viewed at wileyonlinelibrary.com]

Figure B3 shows where a random sample of observations C_i distributed evenly across the range of observed values falls within the posterior predictive distribution of $\tilde{C}_i$. Note that the observations and distributions presented are in the outcome scale, not the probit scale. This reflects both uncertainty in the central estimate $\hat{C}_i$ and simulated draws from the estimated distribution of the residual error. The posterior predictive distributions are consistent with the observations, but without evidence of overfitting. In particular, there is no suggestion that the quality of the fit varies systematically across level, or based on the age at which completers were observed (validating the specification of the age profile). Note that the tail behaviour of the plots, particularly the primary plot, is an extension of the probit transformation used by the model. Specifically,

TABLE B1 Pareto-k diagnostics

Level	Sex	$k > 0.7$	$k > 0.7$ (excluding known points)
Primary	Female	0.019	0.006
Primary	Male	0.018	0.006
Primary	Total	0.021	0.007
Lower secondary	Female	0.026	0.011
Lower secondary	Male	0.024	0.012
Lower secondary	Total	0.031	0.014
Upper secondary	Female	0.032	0.018
Upper secondary	Male	0.031	0.016
Upper secondary	Total	0.035	0.018

when generating a normal replication in the probit space and subsequently transforming the replications back to the original space, the variability in the extremes compresses as present in the plots.

The LOO-CV exercise found that the model seems reasonably well-calibrated (Gabry et al., 2019; Vehtari, Gelman, Gabry et al., 2020; Vehtari et al., 2017; Yao et al., 2017). There is some suggestion that the model slightly overdisperses but that is not unexpected when considering the diversity in observed values and trends present across levels. The PSIS diagnostics are summarised in Table B1. The percentage of points with $k > 0.7$ is fairly small. As has been the trend, the model seems to perform best on the primary school data. The PSIS diagnostic tends to identify points that are outlying or possibly influential. Given that certain model parameters, notably age-reporting and late completion, rely on select few points for estimation, we expect there to be a baseline number of influential points picked up by this test. Concretely, for a country with few surveys, the a_3 and a_4 observations entirely determine the late completion parameter and so in a LOO scenario, we expect these points to be flagged. To given a sense of the percentage of outlying points, we remove these points from the list of $k > 0.7$ points and recompute the percentage of influential and outlying points.