WG/HHS/3

ADJUSTED BAYESIAN COMPLETION RATES (ABC) ESTIMATION TECHNICAL REPORT

Adjusted Bayesian Completion Rates (ABC) Estimation Technical Report

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

[1] Global Education Monitoring Report (GEMR) - UNESCO
[2] University of Toronto
[3] University of Massachusetts Amherst

27 October 2020

Contents

1 Introduction

Sustainable Development Goal 4 on education and its monitoring framework have shifted attention away from mere enrolment towards school completion and learning. SDG target 4.1 on education completion and target 4.5 on equal access have increased interest in using household survey or census data to monitor education indicators. Indeed, there is no alternative for some indicators, such as completion rates for specific population groups, since administrative data on graduates by age or other population characteristics are rarely available.

Survey data bring their own challenges. Most surveys are conducted every three to five years and the results released at least one year later, generating a considerable time lag.

For several indicators, multiple surveys are available. These may provide conflicting information. The 2016 Global Education Monitoring (GEM) Report raised the question of reconciling the different sources (UNESCO, 2016, Box 14.2). 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. It is clear that an assessment of the trend has to consider the relatively larger uncertainty of the 2003 estimate, and that recent DHS and 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.

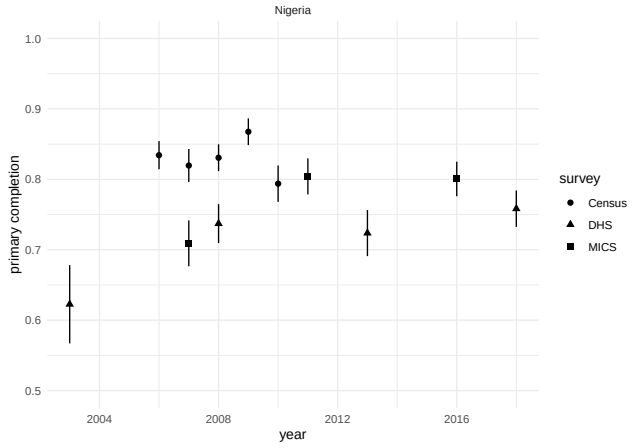


Figure 1: Primary school completion rate in Nigeria five years above nominal age for final primary grade, from different surveys, with estimated 95% uncertainty intervals. N.B. y-axis starts at 0.5.

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 a similar challenge in measuring indicators, such as under 5 mortality or maternal mortality rates, based on multiple sources. The UN Inter-agency Group for Child Mortality Estimation adopted a consensus model to generate annual estimates for under 5 (Alkema and New, 2012) and neo-natal mortality (Alexander and Alkema, 2018) in each member state. The Inter-Agency Group for Maternal Mortality Rates followed a similar process (Alkema et al., 2016).

Here, we introduce a model that builds on these models for health indicators, but is fully adapted to estimating school completion rates. As a result, it is simpler in some respects, but more complex in others. We present the structure of the model and how it addresses a number of specific challenges arising in the context of the

completion rate indicator, present a first set of estimates and demonstrate their robustness and superiority over more simplistic approaches.

This model allows us to estimate and project completion rates for countries and regional aggregates that are less sensitive to individual surveys, in which years they are conducted, and which survey happens to be the latest available for a given country. Figure 2 illustrates some of the results by income group. Specifically, what is shown is both the estimated 'ultimate' rate of school completion and the rate for the age group underlying the official 'Completion Rate' indicator.

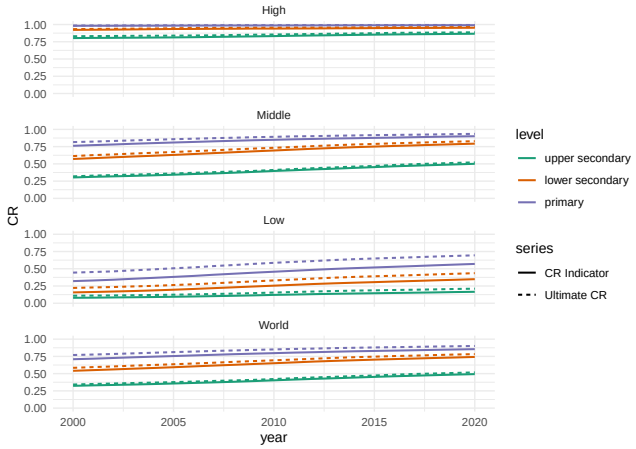


Figure 2: Population-weighted income group averages of projected completion by level.

The analyses in this paper are based on a consolidated collection of 551 microlevel datasets on school completion from 154 countries. The sources are described in detail in Section 6.1.

2 Modelling school completion rates

2.1 School Completion Rates

The agreed SDG indicator of school completion measures completion among individuals who are between three and five years above the theoretical age for the final grade of the education level in question [**REF**]. 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_l, with 'l' standing for 'last grade'. For brevity, we refer to age $a_l + 3$ as a_3, $a_l + 5$ as a_5, and generally to $a_l + n$ as a_n.

Ideally, the most timely observation of completion would be based on individuals one year above this age. The reason for averaging over three ages instead is to smooth out variation resulting from the potentially small sample size of any given birth cohort in household survey data. The reason for shifting the age bracket up by two years is 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 indicators, 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.

As a general term, 'completion rate' will refer to the average (weighted) proportion of a given set of individuals who have completed the level of schooling in question. In the following, $C_{a,c,y}$ is the observed average completion at single-year age a (as opposed to a multi-year age range) in a given country c in year y, such as 15-year-olds in Nigeria in 2010. As primary, lower secondary, and upper secondary school levels are modeled independently, the level index has been omitted for simplicity.

2.2 The model

Our model for school completion serves the purpose of consolidating estimates from different surveys, provides estimates for years without a survey, and allows for short-term 'now-casts' of current completion rates, as well as possibly for longer-term projections of up to at most 10 to 15 years.

2.2.1 Core model for the underlying trend in completion

Across different levels of schooling and countries, observed completion rates cover practically the entire possible range from 0% to 100%. This points to a first modification our model requires compared to models for mortality rates, where even tragically high rates are small in numerical terms.

To account for outcomes constrained to, but spread across the entire $[0, 1]$ interval, we model $K_{a,c,y} = \Phi^{-1}(C_{a,c,y})$, with Φ the cumulative density function (CDF) of a standard normal distribution. In other words, we model the 'probits' of completion.

Without loss of generality, we parametrise the model in terms of the top of the relevant age interval for the Completion Rate indicator, i.e. we take a_5, as the reference age. Let $\kappa_{a,c,y} = \Phi^{-1}(\Gamma_{a,c,y})$ refer to the unknown *true* completion rate at the outcome scale ($\Gamma_{a,c,y}$) and transformed scale ($\kappa_{a,b,c}$) respectively. The aim is a specification that is parsimonious but flexible.

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 $\kappa_{a_5,c,y}$ that meets the above requirements while remaining relatively simple is an ARIMA(0, 1, 0) process, in other words, in terms of a 'random walk with drift':

$$\Delta\kappa_{a_5,c,y} = \kappa_{a_5,c,y} - \kappa_{a_5,c,y-1} = \gamma_c + \epsilon_{c,y},$$

where

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

The long-term drift γ_c, expected to be positive in general, implies an eventual convergence to 100% completion, including for upper secondary. This is the SDG target, and empirically, educational expansion has been the rule over the long term. In any case, the model is not intended for projections beyond one school generation, i.e. 10 to 15 years.

The above specification does suggest a slightly more complex model in which the ϵ terms are autocorrelated to reflect the possibility of multi-year educational development enablers. However, in the interest of selecting a parsimonious model that avoids identifiability concerns, the 'random walk with drift' model is preferred.

2.2.2 Retrospective estimates and their error

Nationally representative household surveys are conducted relatively infrequently, necessitating an attempt 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.

One solution is to take into account the completion reported by older cohorts who were at the time of survey outside of the age bracket of the completion indicator. In particular, suppose for simplicity that all those who

do complete school do so by the time they reach the top of the age bracket. Suppose that the age bracket for the completion rate is 14 to 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, however, 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 three years prior, in 2012. In this way, a single survey contributes completion rate estimates for a series of years, as illustrated in Figure 3.

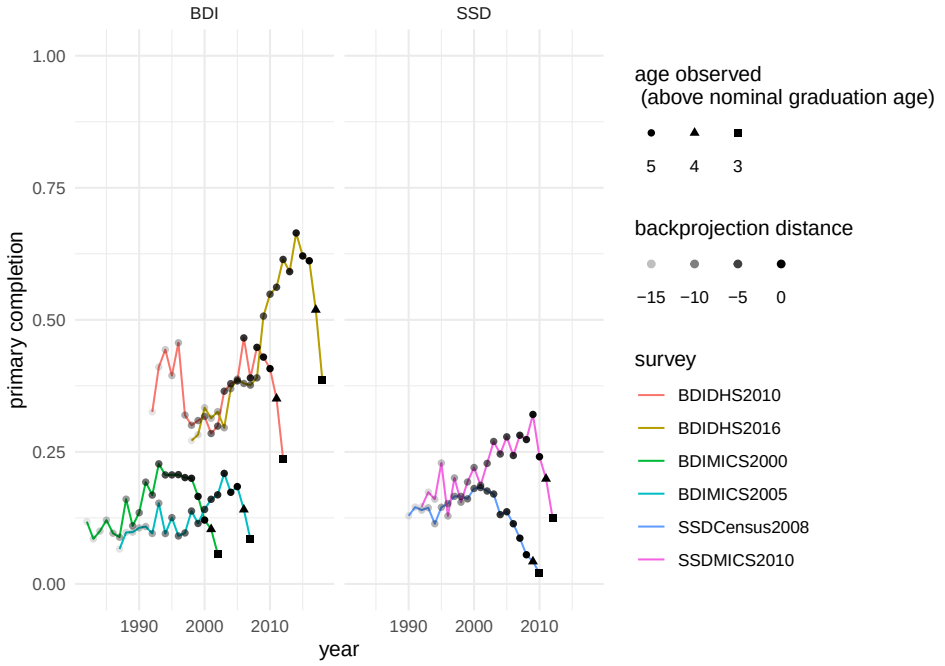


Figure 3: Retrospective series of completion from different surveys. Observations are aligned with the time axis according to the year in which they are at age five years above the nominal age for the final grade. Triangles and squares indicate observations made at younger ages. Faintness indicates retrospective observations of older cohorts.

In general, $C_{a+x,c,y} \approx C_{a,c,y-x}$. One way in which this correspondence fails is if there is if many individuals complete school between the ages of a and $a+x$. This is clear intuitively: primary school completion among 15-year-olds in 2010 is *not* informative of primary school completion among 5-year-olds in 2000, for example. For the time being, assume that all completion of a given level of schooling occurs by the time individuals reach age $a\prime$, and that $a \geq a\prime$. In other words, we consider retrospective estimates that are not distorted by additional completion in the meantime.

Even in this case, the correspondence between $C_{a+x,c,y}$ and $C_{a,c,y-x}$ may not be perfect in general, i.e. observed completion of a given cohort may change at different ages. In particular, completers and non-completers may have systematically different mortality and migration rates. To limit the effect of mortality, we gather retrospective completion rates at a_5 (the top of the indicator age bracket) for individuals aged up to 20 years above the nominal age bracket (i.e. up to a_{25}). In countries with high HIV/AIDS prevalence, even young adult mortality may be relatively high, but in general the largest selection effect in this age range is expected to be differential migration. 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 nonlinear 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 these effects, we assume that the retrospective estimation error increases linearly with age. The adequacy of this specification is investigated further in Section 3.1.1 [posterior predictive fits].

Where multiple surveys are not too far apart, many years will have estimates from more than one survey, as is the case in Figure 3. These overlapping series of completion rates from retrospective estimates 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. In other words, some surveys give systematically higher or lower estimates of completion than others.

2.2.3 Survey bias

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 unbiased, and another biased, but we don't know which is which, then 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 subsample, for instance), or by comparison with a 'gold standard' data source that is assumed to suffer a low bias.

In health applications, some countries possess comprehensive vital registration (VR) systems that can serve as a benchmark. In a functioning VR system, vital events are recorded for the entire population, including *inter alia* births, infant, and maternal deaths. Comparing infant mortality rates (IMR) from a VR system to estimates based on a survey provides some information on the survey's bias. For maternal mortality rates (MMR), VR tends to be biased also because of misclassification of cause of death. In that case, the benchmark estimates are provided by specialised studies. If these estimates of bias are consistent across multiple cases, they determine our expectation of the bias of any given survey of the same type.

An alternative benchmark is provided by censuses. While these share some sources of bias with surveys, specifically the reliance on accurate responses, the sampling frame of census subsamples promises to be more complete than that of surveys.

In the present case of school completion rates, no equivalent to the 'gold standard' of a complete vital registration system or specialised in-depth studies exists. With respect to censuses, a problem shared with similar models for mortality indicators is that data from robust censuses is accessible for only a few of the countries that run DHS or MICS surveys. This is no coincidence, since these surveys were partly motivated by the need to fill an information gap in the absence of high quality census data.

More importantly, the fact that even censuses may miss some subgroups, such as street children, is likely to be more consequential for the estimation of school completion rates than for infant and maternal mortality indicators. For one, childbirth is likely to be rare for some of these groups, whereas school completion rates reference the entire population of a certain age. Moreover, differences in school completion between included and excluded groups are potentially more extreme. It is entirely possibly for primary completion to be almost universal among population in households, but close to zero for among 'missing children'.

In general, a gold standard that allows for an estimation of the *absolute* bias in survey-based estimates is not available in the case of completion rates, therefore. Nevertheless, modelling the bias of available surveys *relative to each other* allows for an unbiased estimation of what would be estimated if surveys of all type were available, even when only a subset or only a single source is. In other words, if series A were consistently lower than series B, then even if for a given year only series A is available, we may still conclude that this is likely to be an underestimate, and that the model estimate should be higher.

We thus model survey s to suffer a bias β_s, and specify the relative structure discussed.

2.2.4 Sampling variance

Survey series and individual surveys differ not only in terms of bias, but also with respect to sample size and sampling variance.

The estimation operates on average completion rates by age as inputs, not on individual-level micro-data. Accordingly, in order to take difference in sampling variation between different surveys (and different age groups) into account, these have to be estimated a priori and provided as input. Not least, this is

computationally more efficient, since these sampling variances only need to be estimated once, and not for each model run or variation in specification.

Some survey reports provide sampling error estimates for selected key indicators, however none do so for school completion rates as defined here. Accordingly, all sampling errors have been estimated from the micro-data, applying the clustered Jackknife procedure that is used to generate the published DHS standard error estimates for other indicators.

In addition to sampling error variance, we assume that an observation i of $\kappa_{a,c,y}$ from a specific survey s is subject to an independent *non-sampling* error. These additional error variance terms account for effects such as variation in enumerator quality or data quality between clusters, implementation flaws, data entry error and so on. It is this non-sampling error that is scaled linearly with the retrospective estimation distance to reflect the increased uncertainty as time passes.

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 extent to which observations *do* differ with respect to sampling variability across countries, but crucially also across ages, time, and surveys series within countries, is shown in Figure 4 for a country with a typical (Mali) and a wide (Belize) spread of estimated standard errors. 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.

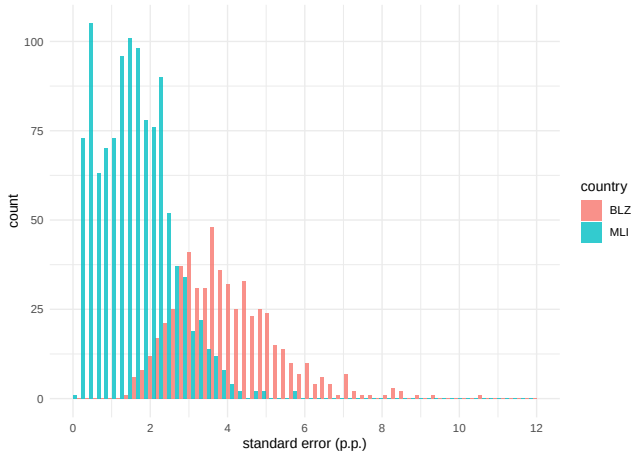


Figure 4: Distribution of estimated sampling standard errors for age-specific completion rates, example countries of Belize and Mali.

2.2.5 Age-misreporting distortion

It is well known that in developing country settings, respondents' ages may be misreported, leading to an overrepresentation of ages that are multiples of five. It is also known that this behaviour correlates with low numeracy skills. Accordingly, it is plausible that age misreporting is likewise associated with school completion.

Indeed, we see clear evidence of this in Figure 5, 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

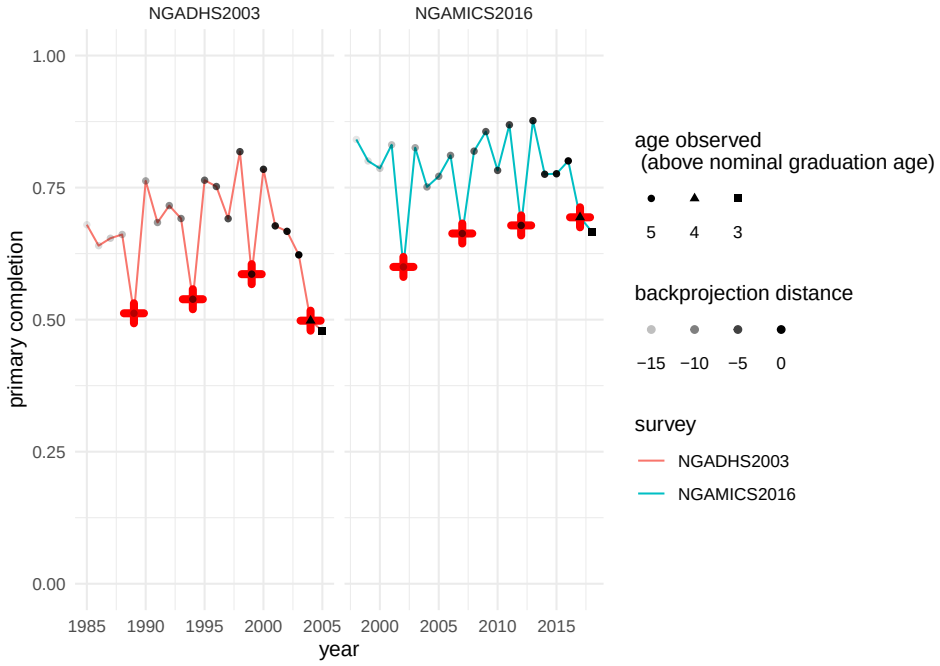


Figure 5: Observed and backcast values of age-specific primary completion in Nigeria. Plus signs indicate observations based on respondents reporting their age at time of survey as a multiple of 5.

the model equation that accounts for the potential distortion in country c due to age misreporting. This distortion is parsimoniously modeled as being rare, but potentially large (see Section 5).

A more complex model could be specified with a micro-foundation, i.e. reflecting the pool of less-educated likely to misreport, for instance. However, too little is known to do this convincingly, and it would introduce undue complexity for what is ultimately a nuisance factor. For example, if the prevalence of age-misreporting were a function of primary completion, the estimation of completion rates at other levels could no longer be done independently, but completion at all levels of schooling would have to be estimated jointly, which is computationally unfeasible. 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. Accordingly, the offset is not modeled explicitly as affecting specific ages, but is allowed to be implicitly absorbed in the overall country intercept.

2.2.6 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 equals 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.

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 this case it is evident that late completion continues even past a_5.

It is necessary, therefore, to model the age profile of observed completion, in a way that allows for late

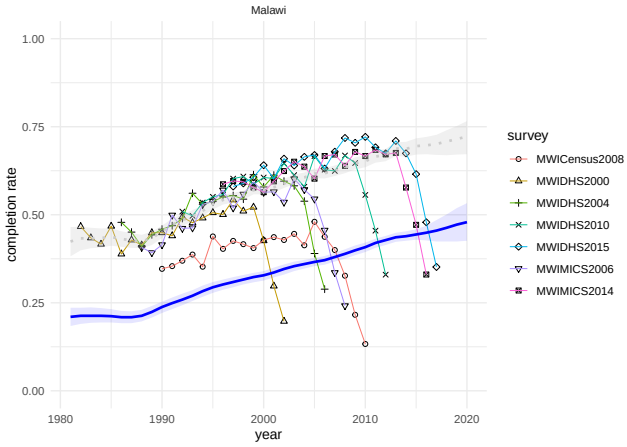


Figure 6: Observed and backcast values of age-specific primary completion in Malawi. The fitted Completion Rate indicator and ultimate completion are indicated by the blue and grey lines respectively.

completion in addition to the error associated with retrospective observations. As a parsimonious but flexible specification, we model the age profile as piece-wise linear in the probit-transformed space with two segments. This is illustrated in Figure 7.

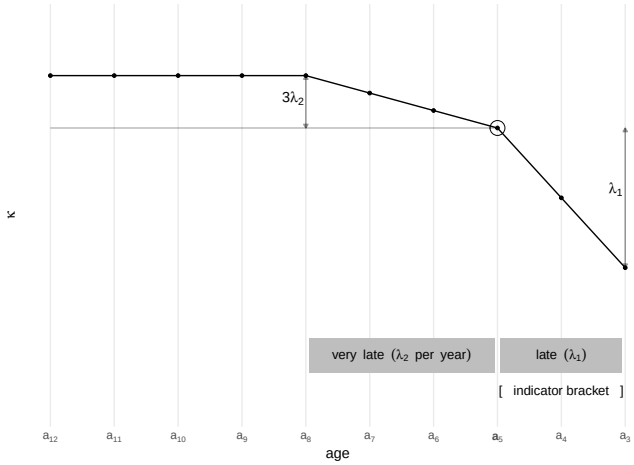


Figure 7: Piece-wise linear model of age-profile of completion.

Specifically, completion between ages a_3 and a_5 is assumed to potentially be lowered by a country-specific constant age slope λ_1 (that is linear in the probit-transformed space), very late completion over the next three years is modelled by the second linear term with country-specific slope λ_2, and further late completion is captured implicitly by the increasing variance as years progress further back.

2.2.7 Population weights

What the above model delivers are estimates of age-specific completion rates $C_{a,c,y}$.

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

$$CR_{c,y} = C_{[a_3,a_5],c,y} = \sum_{i=3}^{5} \frac{p_i}{p_{[3,5]}} C_{a_i,c,y} \ .$$

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 $[a_3, a_5]$.

However, the variation in p_i does not provide useful information regarding school completion.

Recall that the purpose of averaging over several single year age groups in the first place is to smooth out random variation caused by small sample sizes of single year cohorts. Back-of-the-envelope calculations suggest that random variation in the age distribution within the age interval $[a_3, a_5]$ will significantly exceed true differences in birth cohort size in all but the largest surveys and extreme fertility settings.[1]

This can be confirmed empirically, as in Figure 8, 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.

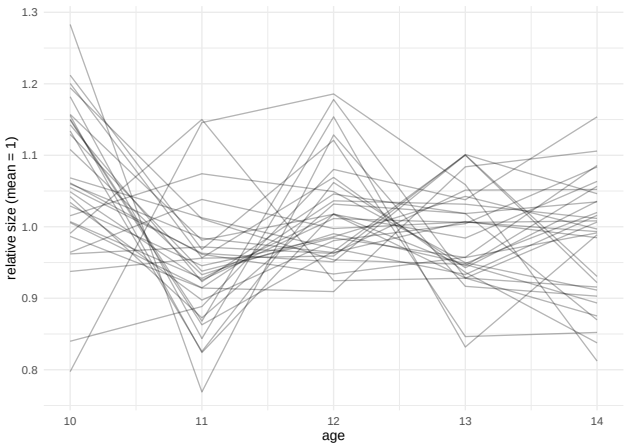


Figure 8: Relative (weighted) size of single-year age groups 10 to 14 in MICS5 surveys.

Finally, recall that one of the purposes of the ABC model is to *project* completion rates. Needing to take into account projected single year cohort sizes would add an extra layer of complexity that would not actually add any insight into the phenomenon of school completion. This would not be helped by using existing population projections. These are typically based on five-year age groups, and where single years of age are projected, these are often model-based, smoothed derivations from five-year aggregates.

The unweighted average is, therefore, arguably preferable in general, and is certainly more suitable for modelling and projection. This age-standardized 'Completion Rate' indicator $CR^*_{c,y}$ is:

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

While this approach is preferable in terms of estimation and understanding the true trend in education system performance, this means that our estimated completion rates may differ marginally from those customarily reported based on individual surveys.

[1] Supposing 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%.

In practice, we are of course interested in the corresponding average of the *true* completion at these ages, i.e. in the results our 'estimated completion rates' are:

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

Also of interest is the 'ultimate cohort completion', which in our specification is assumed to be reached within 8 years after the nominal age for the final grade and is therefore proxied by $\kappa_{a_8,c,y} = \kappa_{a_5,c,y} + \phi_{a_8,c}$.

3 Example 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.

3.1 Fit

3.1.1 Posterior predictive fit

Figure 10 shows where a random sample of observations C_i distributed evenly across countries fall within the posterior predictive distribution of $\tilde{C}_i$. This reflects both uncertainty in the central estimate $\hat{C}_i$ and simulated draws from the estimated distribution of the residual error.

Posterior predictive checks should not be used for formal model selection because the same data are used to estimate the model and assess its fit. In the next section, we therefore assess the predictive performance of the model on out-of-sample observations. Nevertheless, the posterior predictive distributions across all observations provides a first sense of the overall appropriateness and plausibility of the model specification and highlight possible problems.

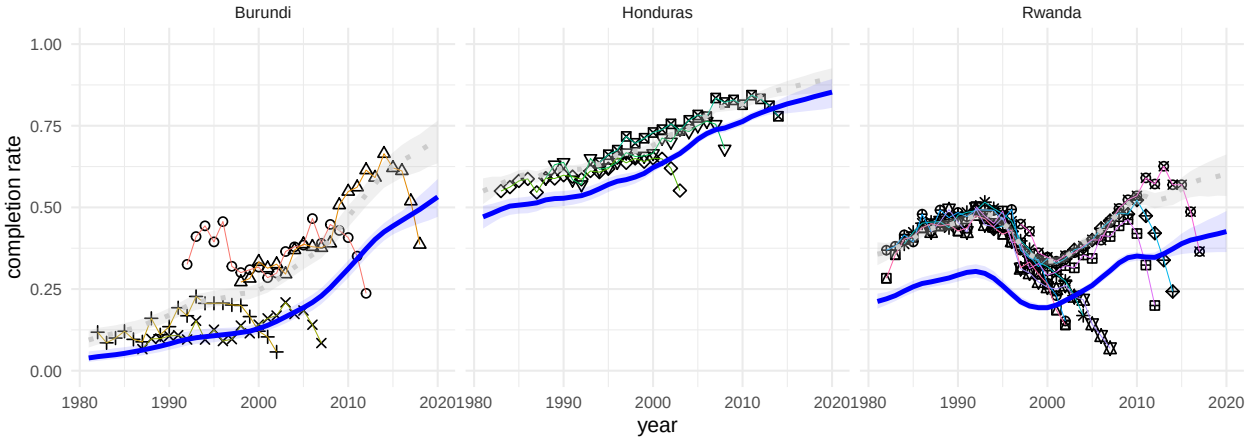
It is evident that the model seems well-calibrated. The posterior predictive distributions are consistent with the observations, but without evidence of overfitting. In particular, there is no evidence that the quality of the fit varies systematically across different levels of completion, or the age at which completers were observed (validating the specification of the age profile). Note that the tail behaviour of the posterior predictive fit plots, particularly in the primary plot, is a relic of the tail behaviour of the probit function. Specifically, 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.

3.1.2 Out-of-sample predictive performance

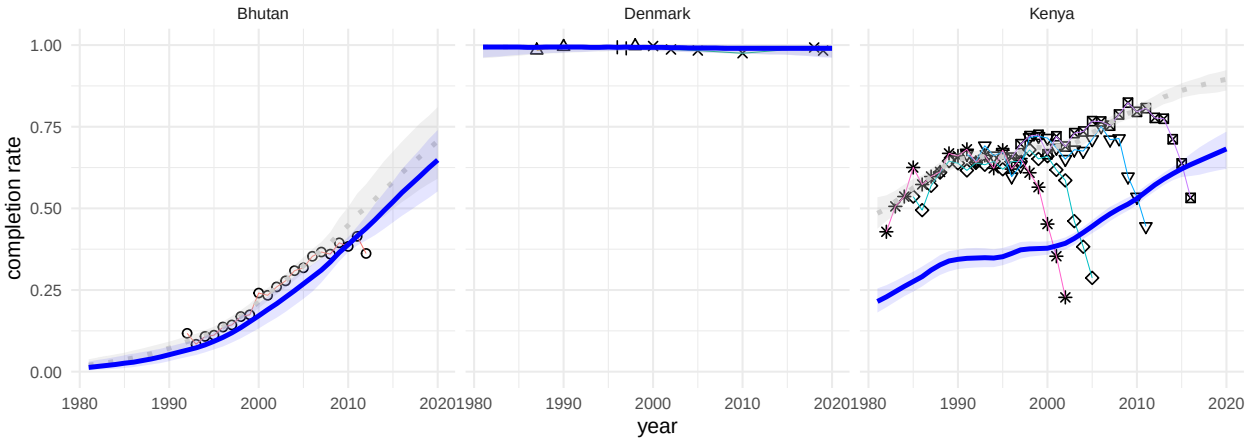
To assess predictive performance, we conduct a 'leave one out' validation. Specifically, all observations based on the latest survey (including backcast values) for each country with more than one survey are omitted from the estimation of the models, and predicted values for these values are obtained.

It is customary to compare predictive performance to benchmark estimates. In the present case, it is not clear what benchmark the ABC model should be compared to by default as there is no previous attempt at modelling the same outcomes using a simpler specification. Here, the ABC model is compared to three alternatives.

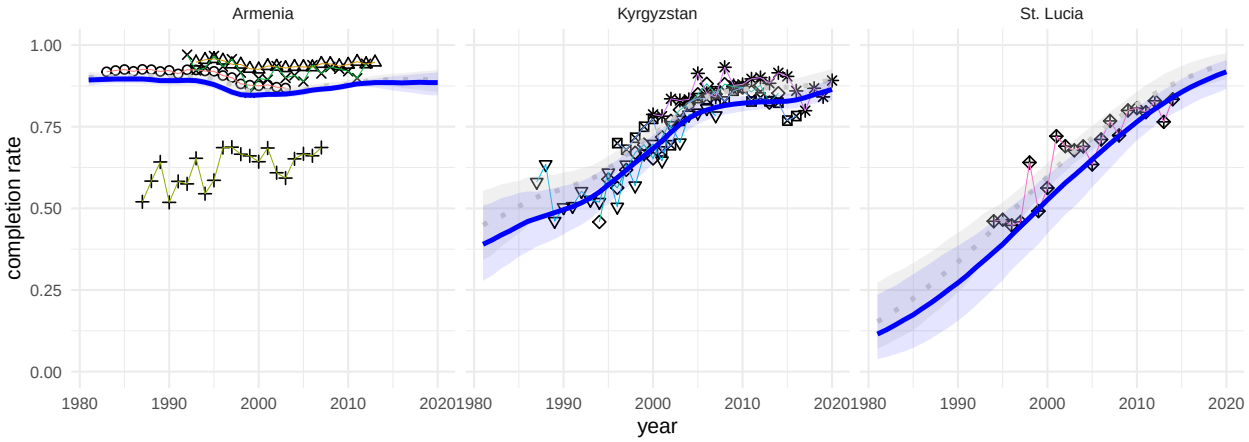
The ABC model is compared to a 'flat' model M^f that simply takes the unweighted mean of all observations as its (invariant) prediction. Formally, it fits an intercept-only least squares model to the κ for a given country. This serves as a naïve baseline.



(a) Primary

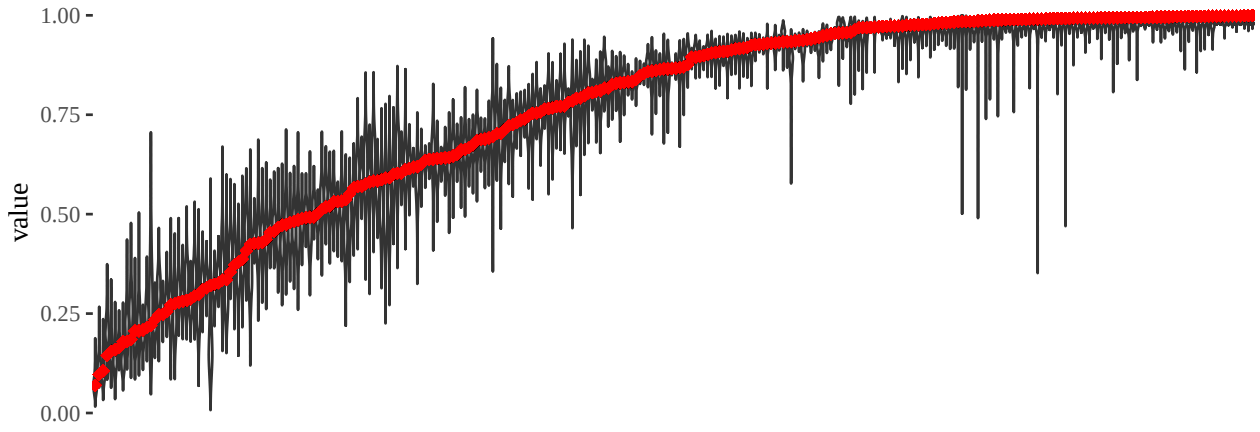


(b) Lower secondary

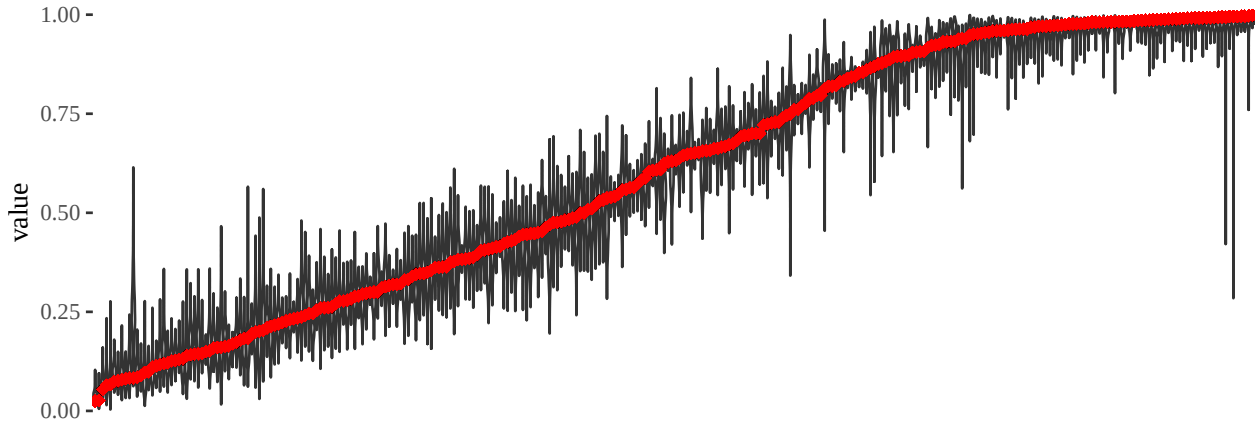


(c) Upper secondary

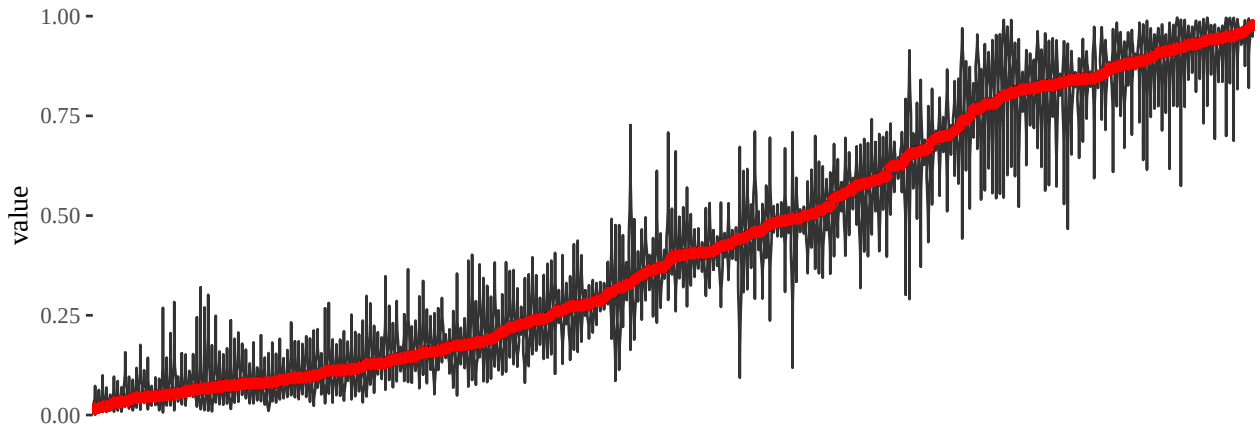
Figure 9: Country projection by level. Solid line = Completion Rate. Dashed line = Ultimate Completion.



(a) Primary



(b) Lower secondary



(c) Upper secondary

Figure 10: Posterior predictive distribution for a random sample of observation.

In addition, a 'simple' statistical model M^s was fitted. This models the κ for a given country as a linear function of an intercept and a slope over time. In other words, 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 other advanced aspects of the ABC model.

First, we compare the prediction error on all the individual C_i arising from the omitted survey. Table 1 shows the results for two measures of prediction error: mean squared error (MSE), as well as mean and median absolute deviations (MAD), all suitably scaled to avoid excessive decimal places. The ABC model offers a meaningful and worthwhile improvement on the simple specifications. This is despite the fact that the bias of the target survey is not exploited, since empirically it turns out the survey family is not informative of the bias to be expected of a given individual survey (see Figure 12) including the omitted one. On the whole, the advantage of the ABC model is higher for lower levels of schooling where effects such as late completion appear more pronounced.

Table 1: Prediction error in age-specific completion rates derived from most recent survey

Model	**Level**	**Sex**	**MSE × 1,000**	**MAD(mean) × 100**	**MAD(median) × 100**
ABC	lsec	female	3.672	4.123	2.728
flat	lsec	female	14.900	8.600	5.600
simple	lsec	female	6.100	5.200	3.300
ABC	lsec	male	4.799	4.853	3.306
flat	lsec	male	11.600	7.900	5.700
simple	lsec	male	8.800	6.500	4.200
ABC	lsec	total	3.516	3.991	2.547
flat	lsec	total	11.800	7.800	5.300
simple	lsec	total	6.500	5.400	3.200
ABC	prim	female	3.097	3.574	2.045
flat	prim	female	14.700	8.300	5.200
simple	prim	female	7.400	5.500	3.100
ABC	prim	male	4.191	4.241	2.482
flat	prim	male	11.300	7.500	4.900
simple	prim	male	11.500	7.000	3.800
ABC	prim	total	2.972	3.446	1.940
flat	prim	total	11.600	7.300	4.500
simple	prim	total	8.200	5.700	2.900
ABC	usec	female	4.472	4.705	3.197
flat	usec	female	11.400	7.800	5.500
simple	usec	female	5.200	5.200	3.600
ABC	usec	male	6.201	5.743	4.243
flat	usec	male	10.200	7.700	6.100
simple	usec	male	7.800	6.600	4.800
ABC	usec	total	4.450	4.766	3.285
flat	usec	total	9.800	7.400	5.500
simple	usec	total	5.500	5.400	3.900

Ultimately, we wish to project not age-specific observations from a given survey, but values of the Completion Rate indicator. We therefore compare in Table 2 how well the predictions $\widehat{\mathrm{CR}}^{\star}$ for different models fit the observed values of $\mathrm{CR}^{\star}$.

Here, another alternative model is included that uses the $\mathrm{CR}^{\star}$ directly instead of aggregating up an estimate from age-specific estimates $C_{a,\cdot,\cdot}$. The 'latest' model M^l simply carries forward the latest observed survey estimate of $\mathrm{CR}^{\star}$ directly. In case there are several estimates available from the same year representing the most recent year apart from the omitted, M^l is equal to their unweighted average. Since many countries only

have a small number of data sources, one of which is omitted moreover, and each survey only contributes a single direct estimate of $\mathrm{CR}^{\star}$, there is no scope for additional statistical modelling on the $\mathrm{CR}^{\star}$ directly. The other models, M^f, M^s, and M^{ABC} of course imply estimates $\widehat{\mathrm{CR}}^{\star}$ by averaging the age-specific estimates.

As before, the ABC model outperforms all the alternatives, and to a greater degree for lower levels of education. Indeed, by some measures, the gain in predictive accuracy from moving from the simple to the ABC model is approximately as large as the gain from moving from a naïve flat average to the simple probit trend.

Note that these statistics ultimately cannot answer the question of which model produces the superior estimates of the *true* underlying completion rate, as measured without error, since there are no known values to compare the estimates to. In particular, for all but the ABC model, the predicted value for the true completion rate and that observed from a given survey coincide. In the ABC model they differ in general, since the latter but not the former potentially includes an estimate of age misreporting and survey bias. Accordingly, the alternative models could in principle generate an estimate that is closer to the observed value measured with error than the ABC estimate of the observed value, but farther from the true value than the ABC estimate of the true value. Ultimately, a conclusive performance assessment can only be conducted in the presence of a reasonable number of 'gold standard' benchmark estimates known to be highly accurate.

Table 2: Prediction error in most recent observed completion rate indicator

Model	Level	Sex	MSE × 1,000	MAD(mean) × 100	MAD(median) × 100
ABC	lsec	female	5.845	5.103	2.954
flat	lsec	female	23.300	10.900	6.800
latest	lsec	female	10.200	6.500	3.900
simple	lsec	female	6.900	5.600	3.300
ABC	lsec	male	6.250	5.376	3.469
flat	lsec	male	12.300	8.000	5.200
latest	lsec	male	7.800	5.700	3.100
simple	lsec	male	6.800	5.700	3.700
ABC	lsec	total	5.595	4.928	3.132
flat	lsec	total	16.300	9.100	5.600
latest	lsec	total	8.500	5.900	3.100
simple	lsec	total	6.500	5.500	3.500
ABC	prim	female	5.341	4.433	1.995
flat	prim	female	18.900	9.300	5.700
latest	prim	female	7.900	5.500	2.400
simple	prim	female	6.500	5.200	2.400
ABC	prim	male	7.879	5.681	2.599
flat	prim	male	9.100	6.500	3.900
latest	prim	male	8.400	5.900	2.900
simple	prim	male	7.300	5.400	2.300
ABC	prim	total	5.834	4.608	2.178
flat	prim	total	12.000	7.500	4.900
latest	prim	total	7.900	5.600	2.800
simple	prim	total	6.300	5.000	2.300
ABC	usec	female	5.379	5.124	3.008
flat	usec	female	19.300	10.800	9.000
latest	usec	female	7.200	5.900	3.700
simple	usec	female	5.800	5.500	3.500
ABC	usec	male	8.236	6.324	4.737
flat	usec	male	13.000	8.700	7.400
latest	usec	male	7.800	6.500	4.800
simple	usec	male	8.700	6.600	4.700
ABC	usec	total	6.042	5.420	3.718
flat	usec	total	15.000	9.400	7.300
latest	usec	total	6.600	5.800	3.800
simple	usec	total	6.000	5.600	3.900

3.2 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 11 summarises the posterior distributions for survey bias, late completion, and age misreporting.

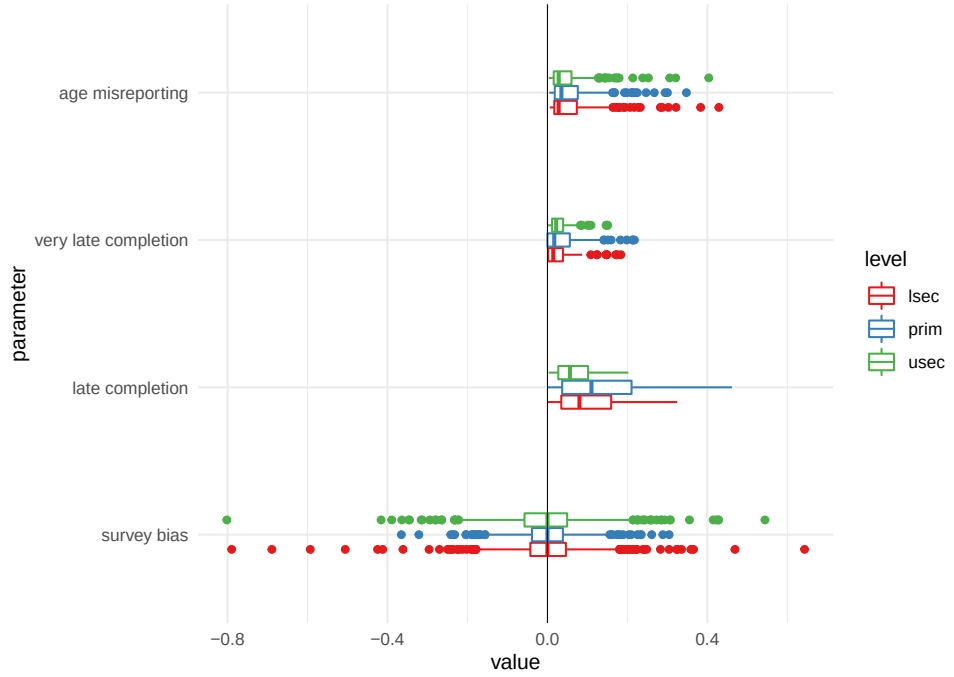


Figure 11: Posterior distribution of ABC model parameters, inverse probit scale.

Recall that the two-stage specification of the age misreporting effect: whether it occurs, and if, then how strongly. Figure 11 displays the second parameter, the strength of the age misreporting effect conditional in those countries where it is detectable. 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 exceeds 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 11 are shown on the scale of κ. 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. Nevertheless, to put the estimated magnitude of late completion into perspective: at a lower secondary completion rate of 50%, a late completion effect of 0.25 on the nonlinear scale of Figure 11 corresponds to a gap of more than 9 percentage points between completion observed at the bottom and the top of the age bracket.

The finding that late completion is lower at higher levels of schooling is interesting. A priori, it might be expected that a longer school career up to that point would have created more opportunities for delays. Also, the next higher level, tertiary education, is completed by many *much* later than the 'theoretical' age for timely completion. Countering these effects, the fact that in our results there is less late completion at the upper secondary level suggests that 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 naturally centre on zero. What was not known a priori is that the distribution of survey bias has 'fat tails', in other words, outliers in the form of distributions that are 'off' by a large amount are fairly common. This calls for

great caution in interpreting 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. We see that the conclusion that some individual surveys suffer heavy bias is confirmed with great confidence. However, it is also apparent that there is no discernible systematic difference between DHS and MICS families, or specific waves: the estimated systematic bias common to surveys of a given type is practically indistinguishable from zero. This implies 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 street children, they tend to be missed by household surveys in general. 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.

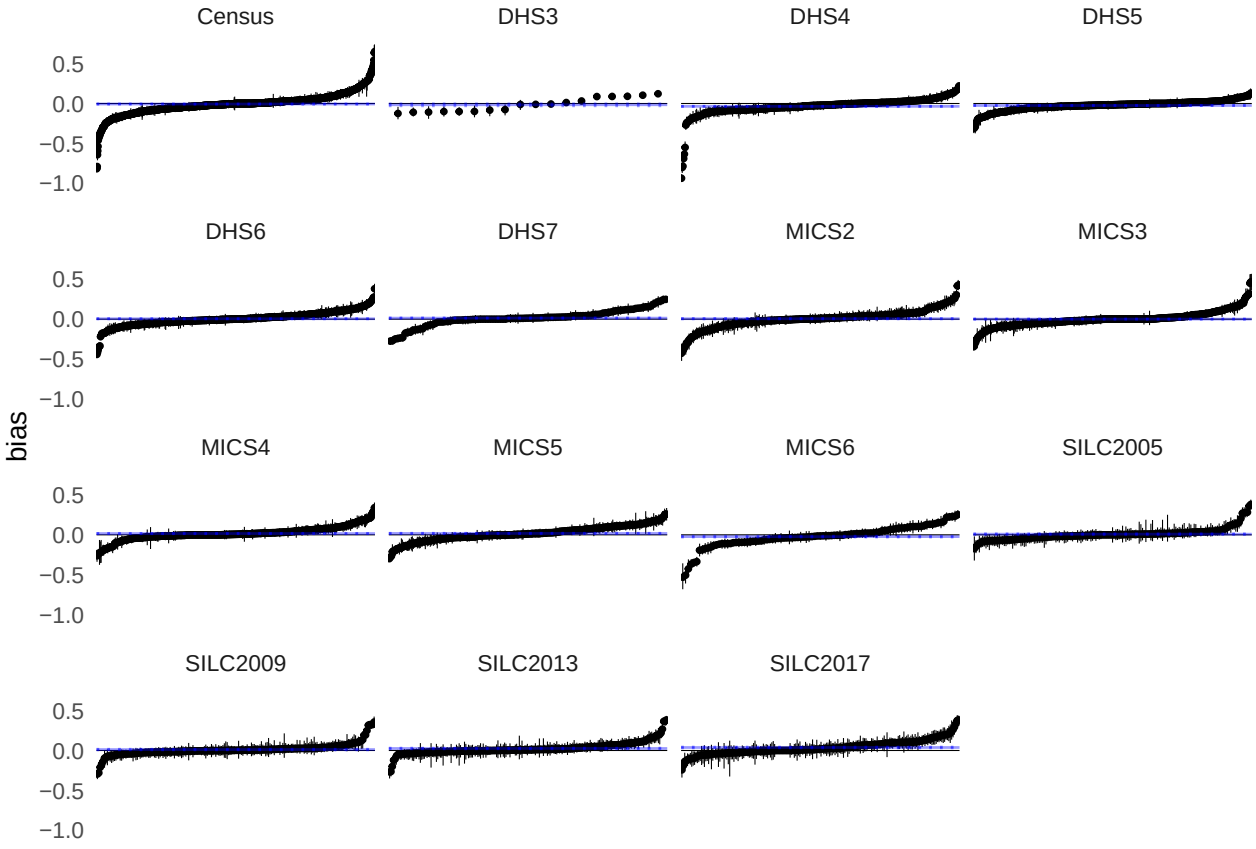


Figure 12: Posterior distribution of individual survey bias terms, inverse probit scale. Blue lines: median (solid), 10th and 90th percentile (dotted).

4 Conclusion

Global model-based estimate of development indicators should not be mistaken for 'the real thing'. This is especially true when they are used to assess the attainment 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 2014. 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 9. That said, this model specification would not explicitly 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 were to occur, an explicit change in long-term drift would not be captured, leaving the impact to a realignment though a shock in the residuals.

Two key challenges are identified. At a fundamental level, absent unbiased 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. Secondly, while the principal rationale for the CR indicator was that individuals in the reference age brackets could be assumed to have completed the school level in question, our results show that severe amounts of very late completion, even five 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 atypical late completion patterns.

At some level the finding of such widespread and considerable late completion is disappointing, 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 *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 statistical negligible. The age bracket 3 to 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 can *not* be interpreted as a proxy for ultimate cohort completion. Instead, it should be recognised as measuring what might be termed 'reasonably timely completion'.

5 Annex: Full model summary

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 $\kappa_{a[i],c[i],y[i]}$, 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^2 and ω^2 respectively:

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

where the underlying 'true' values follow:

$$\Delta\kappa_{a_5,c,y} = \kappa_{a_5,c,y} - \kappa_{a_5,c,y-1} = \gamma_c + \epsilon_{c,y},$$

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.

We assume:

$$\begin{aligned} \epsilon_{c,y} &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ \sigma_\epsilon &\sim \text{Gamma}(2, 0.1) \end{aligned}$$

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 completing, note that for $C = 0.5$, i.e. 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 0.035, an entirely plausible value for the jitter term while still allowing for reasonably large shocks. Returning to the discussion on the comparison of the transformed and outcome spaces, consider the tail behaviour of the probit curve. If we observe 99% completion, a 0.5 percentage point down to 98% translates to a change in κ of approximately 0.16. A 0.5 percentage point increase translates to a change in κ of 0.25 whereas a 1 percentage point increase diverges. If we were to observe 98% completion, a 0.5 percentage point down to 97.5% translates to a change in κ of approximately 0.09, a dramatically smaller difference as compared to the 99% case. Recognizing that noise in the extreme values in the outcome space may appear as more than simply noise due to the drastic scaling difference, we impose a cap on extreme values such that if a country observes its maximum value above 98% completion or minimum value below 2% completion, all of its observations are uniformly shifted inwards such that the maximum is now 98% or minimum is 2% respectively. This reduces the risk of noise having undue influence on drift while retaining the outcome space drift structure. In post-processing, the extracted true values are shifted back to restore the original levels.

With respect to the country-specific drift γ_c, we specify an exponentially-modified Gaussian, specifically:

$$\gamma_c \sim \text{EMG}(\mu = 0, \sigma = 0.025, \lambda = 2).$$

This distribution is shown in Figure 13 to illustrate the motivation behind it. The strongly-skewed character of the distribution is preferred for the drift parameter γ over a symmetrical prior, to require strong empirical evidence before accepting a negative long-term trend. This reflects the fact that logically, a negative trend cannot in fact be the long-term historical experience of a country that is in fact currently far from zero. In addition, for projections up to 2030, the continuation of a long-term negative trend does not meet a priori plausibility criteria as a 'central' (rather than 'low') scenario. This is especially true if a country actually experienced progress in completion most recently.

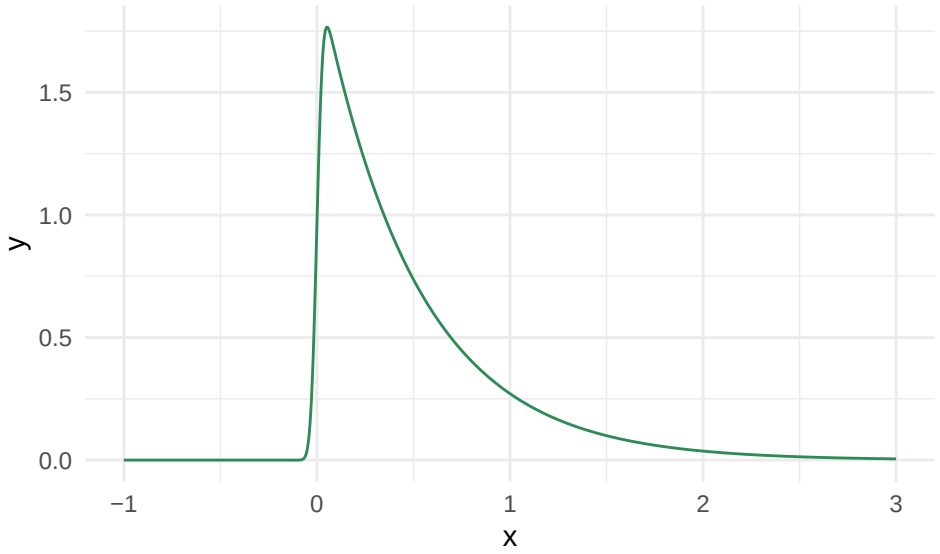


Figure 13: Drift Prior

Despite being quasi-informative in terms of discouraging negative estimates for γ, overall this prior is over-dispersed with respect to γ and leaves its estimates under-determined. To demonstrate this, Figure 14 superimposes the posterior distributions of random γ_c from every quintile, i.e. from the 20% of countries with

the smallest median γ_c, ..., and the 20% of countries with the largest median γ_c. Evidently, these posteriors are much more concentrated than the prior, reflecting the dominant influence of the data. Note that the scale of the y-axis has undergone a square root transformation to assist with visual clarity.

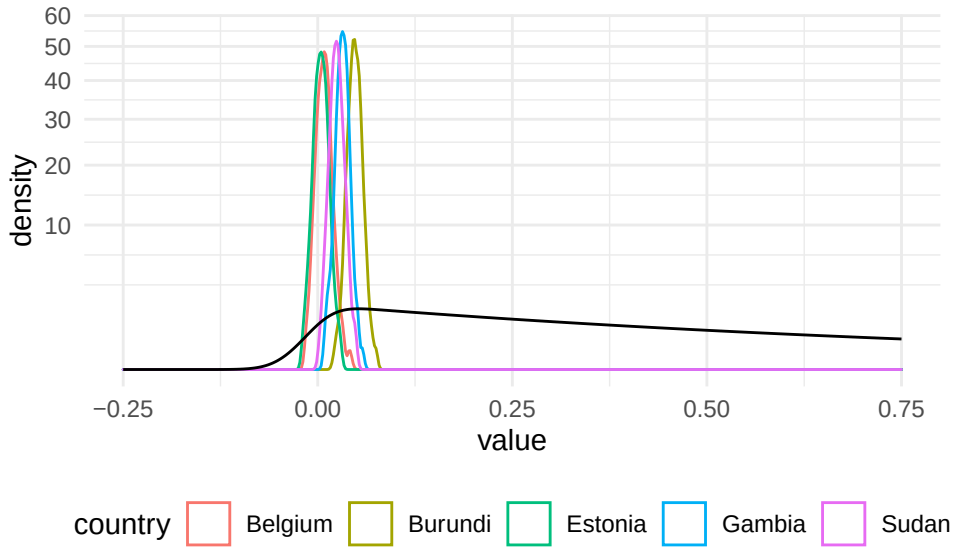


Figure 14: Gamma Prior and Posterior Comparison

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

$$\phi_{a,c} = \begin{cases} 0.5 \cdot (a - a_5) \cdot \lambda_{1c} \cdot \mathbb{1}_l & \text{if } a \in \{a_3, a_4\} \\ \min(3, a - a_5) \cdot \lambda_{2c} \cdot \mathbb{1}_{vl} & \text{if } a > a_5 \end{cases}$$

Priors of $\lambda_{1c} \sim \mathcal{N}(0, 0.25)$ and $\lambda_{2c} \sim \mathcal{N}(0, 0.25)$ correspond to a large share of countries with limited lateness but a significant possibility of late completion of a considerable degree. The indicators reflect the reality that late completion is only estimated if there is an indication of its presence being more than simply noise. In the case of λ_{2c}, it is estimated for those countries with median observed values below 0.95. In the case of λ_{1c}, it is estimated for those countries with median observed values below 0.95 *or* $[a_3, a_5]$ values consistently below $[a_5, a_7]$. 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.

Survey bias β_s has the following prior distribution:

$$\begin{aligned} \beta_s^* &\sim \text{Cauchy}(0, \sigma_{\text{bias}}^2) \\ \sigma_{\text{bias}} &\sim \mathcal{N}(0, 0.25) \end{aligned}$$

The Cauchy distribution is 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 in Figure 9.

Recall that the relative structure of the survey bias induces a sum-to-zero behaviour in the survey bias terms for each country. Equivalently, there is a redundancy in an unconstrained relative survey bias specification that produces highly undesirable geometry in the Bayesian framework. To remedy this fault, we explicitly specify a sum-to-zero constraint on a per country basis to reduce the degrees of freedom. That is, for a country with n surveys, $\beta_{sj} = \beta_{sj}^*$ for surveys $j = 1, ..., n-1$, and $\beta_{sn} = -\sum_{i=1}^{n-1} \beta_{si}^*$ for survey n. We acknowledge that such a specification overdisperses the implied prior on survey n, however, empirically, different permutations of survey ordering did not produce materially different results. Further, the transformations required to produce symmetric priors for all surveys after the constraint proved to be unfeasible with the Cauchy distribution, which is critical for modelling outliers.

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

$$\tau_c \sim \text{Exp}(10)$$

The sampling variance ν_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 transformed completion rate $K_{a,y,c,s}$ in year y at age a in country y from survey s is estimated as (omitting indices for clarity):

$$\widehat{Var}(K) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (K_i - K)^2$$

where

$$K_i = nK - (n-1)K_{(i)}.$$

Here, K is calculated on the full sample, $K_{(i)}$ is calculated on the sample with the i^{th} cluster excluded, and n is the total number of clusters. 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 $1,000$ 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 as $\nu_i^2 = \frac{\widehat{Var}(K)}{(f(\Phi^{-1}(K)))^2}$ where f and Φ^{-1} are the density and inverse CDF of the standard normal distribution respectively.

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:

$$\omega_{a,s}^2 = (1 + 0.05 \cdot \max(0, a - a_5)) \cdot \omega_s^2$$
$$\omega_s \sim \text{Gamma}(2, 4)$$

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

6 Implementation

6.1 Data

The analysis is based on a consolidated collection of individual-level microdata on school completion. The results presented here are based on 551 distinct surveys from 154 countries.

Specifically, sources include Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS) and selected other household surveys that form the basis for the Global Education Monitoring Report's *World Inequality Database on Education* (WIDE). Because WIDE focuses on the analysis of inequality, it does not include census data, such as the public-use samples made available and popularized for research by IPUMS.

While (binary) gender information is collected in censuses as a matter of course, and urban/rural location is very common, but by no means universal, there is no census equivalent of the DHS/MICS wealth index that allows for the analysis of socioeconomic inequality between households in different quintiles of the wealth distribution. However, for the present analysis, census samples were added that contain the necessary information on school completion. For computational reasons, these IPUMS extracts were limited to 1 million observations each.

6.2 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-apple-darwin17.0 (64-bit) platform. The present exercise draws inspiration from the `distortr` package[2] by Monica Alexander that underpins the similarly-motivated models for infant and maternal mortality.

4 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 estimation for the primary completion rates model for females and males together ran for 1801 minutes. This runtime reflects sequential processing of all chains though multiple cores were used to run the models by level, sex, and chain in parallel. Parallel computation was executed using the `future` and `clustermq` packages[3][4] and the `drake` package[5] was used to ensure reproducibility and assist with version control.

6.3 Convergence

Convergence was assessed both through Gelman's criterion for *potential scale reduction factors* (PSRF) and visual inspection of traceplots for all parameters. Figure 15 displays the PSRF values of the 9 parameters with the highest (= worst) PSRFs for the estimations for each level and total, female, or male rates. Values below 1.1 are considered acceptable. Example traceplots for these 'worst' parameters from the model for primary education and both sexes are shown in Figure 16. Recall that as described in Section 5, the survey bias has a Cauchy prior which is implemented through the tangent transformation of a $Unif(-\frac{\pi}{2}, \frac{\pi}{2})$ distribution, thus explaining the upper bound in the respective traceplot.

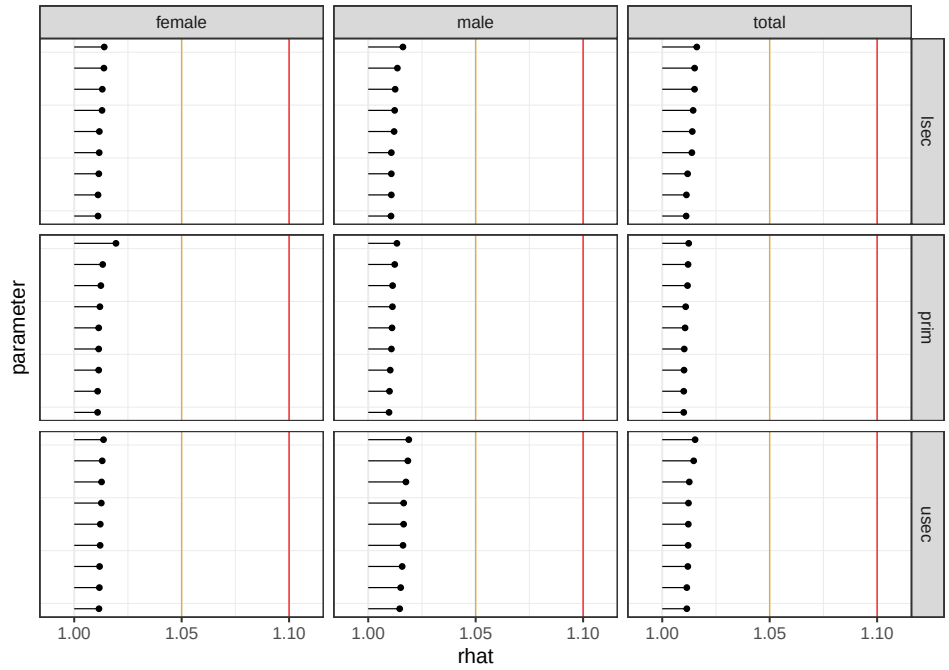


Figure 15: Highest PSRFs by Level and Sex

[2]https://github.com/MJAlexander/distortr
[3]https://github.com/HenrikBengtsson/future
[4]https://github.com/mschubert/clustermq
[5]https://github.com/ropensci/drake

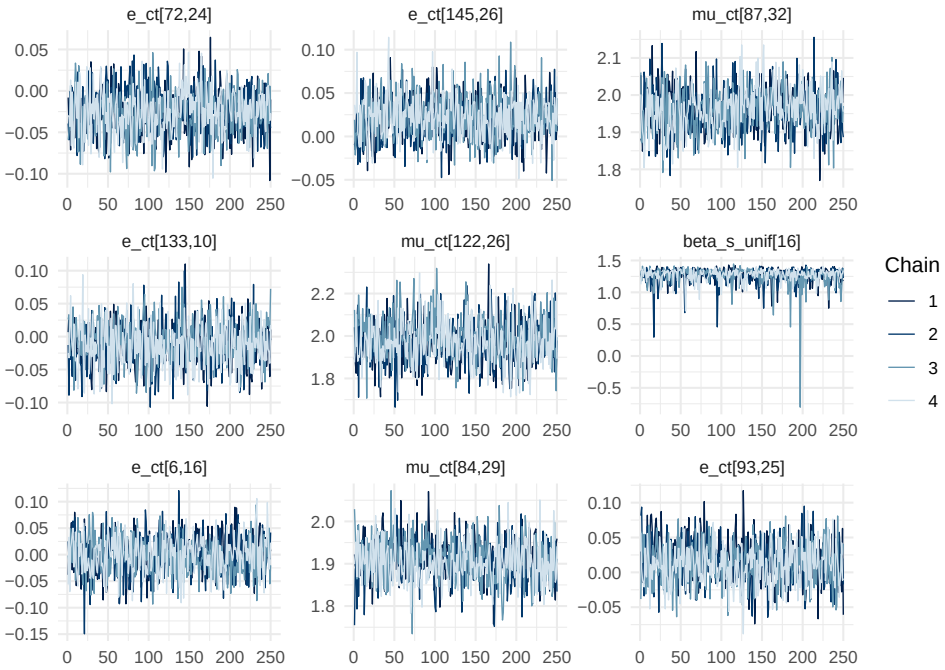


Figure 16: Example Traceplots