United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

TECHNICAL COOPERATION GROUP

SUSTAINABLE DEVELOPMENT GOALS

A new methodology for estimating completion rates

Global Education Monitoring Report

TCG Fifth Meeting
15-16 November 2018
Mexico City, Mexico
TCG5/ REF/7

Introduction

The combination of SDG targets 4.1 on completion and 4.5 on equity have increased interest in the use of household survey or census data for monitoring. Indeed, unlike during the Education for All period, when the interest was limited to averages, there is no alternative to household survey data, since administrative data by several population characteristics are rarely available. But survey data bring their own challenges. Two stand out.

- Multiple surveys are often available and may provide conflicting information. The 2016 Global Education Monitoring (GEM) Report raised the question of reconciling the different sources (UNESCO, 2016; Box 14.2). Averaging estimates or fitting a trend ignores relevant information. Some sources may systematically result in lower or higher estimates relative to others, reflecting differences in sampling frames or how questions are asked. Some sources may show greater variability due to small sample size or other, non-statistical issues that make them less reliable. Some respondents provide information retrospectively and the time that has lapsed increases the risk of errors that need to be corrected.

- Most surveys are conducted every three to five years and the results released at least one year later, generating a considerable time lag. This lack of up to date is often used as an argument against the use of surveys for reporting on SDGs. Yet, available sources contain sufficient data to base a short-term projection to the current year.

The international health community faced similar challenges in measuring indicators, such as infant, child or maternal mortality rates. As in education, these required data coming from multiple sources, some of which were out of date. 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 country. The Inter-Agency Group for Maternal Mortality Rates followed a similar process (Alkema et al., 2016).

In this paper, a model is presented that builds on and adapts these health indicator models to the estimation of school completion rates. Following a presentation of the structure of the model, the paper discusses how a number of specific challenges arising in the context of the completion rate indicator are addressed. The paper also presents selected estimates and demonstrates their robustness and superiority over more simplistic approaches.

Modeling school completion rates

SDG thematic indicator 4.1.4 on school completion 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. For example, if a child is supposed to enter the final grade of primary school at age 11, the indicator is estimated for the age group of 14-16 year olds.

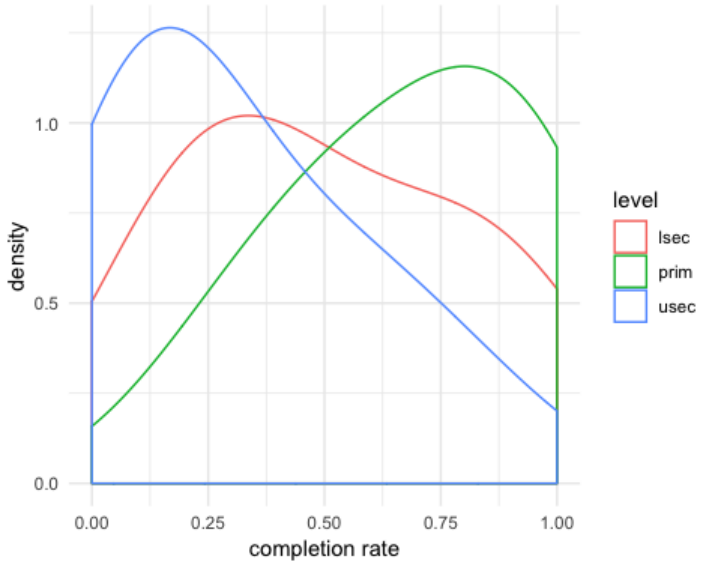
Ideally, completion rates should be observed among individuals one year above this age.

- The reason for averaging over three years is to smooth out variation resulting from the potentially small sample size of any given single-year cohort in household survey data.

- The reason for shifting the age bracket up by three 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* complete school. Accordingly, the completion rate indicators seeks to abstract away from the question of timeliness to some extent and capture all completion that is not unreasonably delayed. In fact, in several countries, a non-negligible share of a particular cohort continues to complete a level of education even more than five years after the theoretical graduation age, an issue that is discussed further in this paper.

Figure 1 shows the distribution of completion rates by country and level observed in Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS). At each of the three education levels (primary, lower and upper secondary), a wide range of completion rates can be observed. This is in contrast to mortality indicators where even high rates are statistically very close to zero.

Figure 1. Kernel density estimate of country-specific completion rates by education level



Terminology

In the following, a 'survey' refers to a single survey sample, e.g. the 2013 Nigeria DHS. By contrast, a given group of surveys sharing a sampling approach and questionnaire refers to a 'survey series', e.g. all phase 5 DHS surveys.

As a general term, 'completion rate' refers to the average (weighted) proportion of a given defined set X of individuals $C(X)$ who have completed the level of schooling in question, which in this paper is always primary school. Such sets are defined by single-year age groups in a given country c in year y, such as 15-year-olds in Nigeria in 2010, $C_{a,c,y}$. The synthetic 'completion rate' as an *indicator* $CR^{*}_{c,y}$ is taken to refer to:

$$CR^*_{c,y} = \frac{1}{3}\sum_{a=a_l+3}^{a_l+5} C_{a,c,y}$$

where a_l is the official entry age into the last grade of the schooling level in question. Note that this definition of $CR^*_{c,y}$ may in general differ slightly from the average completion rate of the three-year age interval $[a_l+3, a_l+5]$. This latter measure, the empirical 'Completion Rate' indicator $CR_{c,y} = C_{a_l+3 \le a \le a_l+5,c,y}$, or more correctly, its estimate $\widehat{CR}_{c,y}$, corresponds to customary published estimates of the 'primary completion rate' based on any given survey, but for reasons discussed below, CR^* is preferred for modelling.

Backcasting

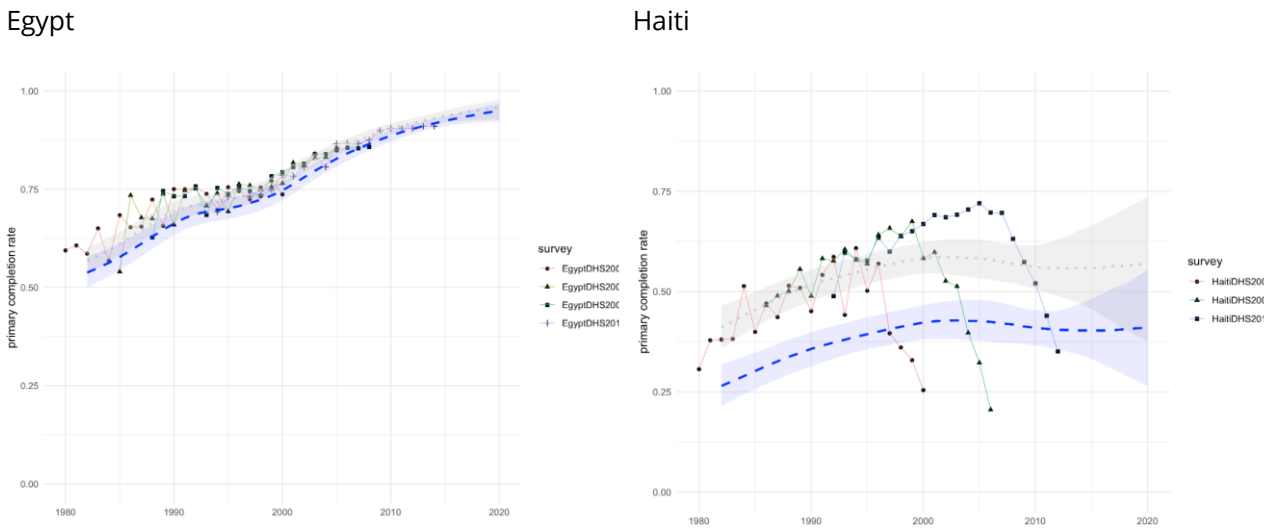
Unlike health-related indicators, such as mortality rates, there are relatively fewer data sources for school completion. If each survey only contributed an indicator estimate for the survey year, for those individuals observed during the nominal age range for the completion rate indicator, many countries would have too few observations to perform any kind of robust statistical trend estimation.

One solution is to 'backcast' completion along cohort lines. In particular, suppose for simplicity that all those who do complete school do so by the time they reach the age bracket that enters into the calculation of the completion rate indicator. Suppose, as before, that the age bracket for the completion rate is 14-16 years 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 three years prior, in 2012. Backcast values are taken as 'past observations', so that a single survey contributes completion rate estimates for a series of years.

Late completion

This correspondence is not perfect and may be assumed to be worse for older cohorts, motivating the specification of a backcasting error terms, which are discussed later. In a number of countries where delays in school entry and progression are severe, even the 'grace period' allowed by the shifted age bracket of 3 to 5 years above the theoretical age for the final grade of the education level in question is not sufficient to ensure that $C_{a_l+3,c,y}$ equals ultimate completion of the cohort in question, $C_{a_l+5,c,y+2}$. In other words, some individuals complete school *during* the age interval $[a_l+3, a_l+5]$.

Figure 2. Kernel density estimate of country-specific completion rates by education level



This is highlighted in **Figure 2** with two contrasting examples. In Egypt, children generally complete primary school during the age bracket specified for the indicator. In Haiti, children complete primary school well after the age bracket, as highlighted by the long 'tails' of the observed completion rate by single age group: according to the 2012 DHS, the completion rate of 14-16 year olds, corresponding to the 2012 cohort, was half that of 19-21 year olds, corresponding to the 2007 cohort.

Strictly speaking the cross-sectional profiles from any given survey do not show the degree of late completion between ages a_{l+3} and a_{l+5}. In principle, they could also arise from a dramatic decline in ultimate completion between successive cohorts. However, we know from contextual information that no such collapse occurred in the case of Haiti, as the same pattern is observed in the earlier DGS rounds in 2006 and 2000.

Late completion poses a problem for the estimation, because backcasting completion observed at some higher age is only valid down to the age after which no more school completion occurs. For example, if it is known, or we are willing to assume, that all those who do complete primary school will do so by age 17 at the latest, or conversely, that no 17-year-old who has not yet completed primary school will still do so, then observed primary school completion status at age 37 is informative of primary school completion status at age 17 twenty years prior. At the aggregate level, this is subject to caveats concerning selective survival and migration, but still holds with respect to transitions in school completion status per se.

However, in the case of Haiti, for instance, since many individuals evidently do still complete school between the ages a_{l+3} and a_{l+5}, and even later, the current completion status of older cohorts can *not* simply be used as backcast past completion status at age a_{l+3} even if migration and mortality are negligible. At best, we can backcast back to past completion at age a_{l+5}.

This situation motivates the specification of the core model (see Annex) in terms of $C_{a_l+5,c,y}$, for which backcast 'observations' are available. However, completion at ages a_{l+3} and a_{l+4} must still be modeled

in order to estimate the completion rate indicator. Graphical inspection suggests that late completion can follow many different age patterns that are not easily summarised.[i]

A disadvantage of this simple specification is that, since it only allows for late completion up to age $a_l + 5$, completion beyond this age will depress the most recent observations, but be included in the backcast observations, inducing a spurious negative trend. This is well illustrated by the case of Haiti but also other countries, such as Colombia, Kenya or Lesotho.[ii]

However, note that numerous countries only benefit from one or two surveys, so a more flexible specification for late completion risks preventing the identification of actual cohort-on-cohort declines.

Survey bias

The process so far can only estimate the *relative* bias of different surveys. If *all* surveys overestimate school completion, for example because they exclude street children, this shared bias cannot be identified without additional assumptions 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, some countries possess comprehensive vital registration systems that can serve as a benchmark. In a functioning vital registration system, vital events are recorded for the entire population, including *inter alia* births, infant, and maternal deaths. Comparing infant or maternal mortality rates from a vital registration system to estimates based on a survey provides some information on a survey's bias. If these estimates of bias are consistent across multiple cases, they can 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, in principle they eliminate, and in practice considerably reduce, sampling variation. While some of this advantage is lost when in practice public-use census *subsamples* are analysed, it remains the case that 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 vital registration system 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, therefore, a gold standard that allows for an estimation of the *absolute* bias is not available in the case of survey-based completion rates. 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 is 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. In the model, the bias terms for a survey series (e.g. MICS 5) are drawn from a shared distribution.[iii]

Backcasting error

Nationally representative household surveys are conducted relatively infrequently, necessitating an attempt to exploit as much information as possible from each round. Here, that specifically means taking into account the education reported by older cohorts who were outside of the age bracket of the completion indicator at the time of the survey. The school completion observed among 20-year-olds in 2015, for example, carries relevant information about, but will not in general be exactly equal to what would have been observed among 15-year-olds in 2010.

There are multiple reasons why observations of the same cohort at different ages would differ systematically, beyond sampling variation. Some individuals may have completed school in the meantime. However, in this case the error would not be induced by the backcasting, but the assumption underlying the specification of the completion rate itself would be violated. In other words, if there are significant numbers of 17 or 18-year-olds still completing primary school, say, then the completion rate would be poorly measured even using only information for individuals who at the time of survey were within the strict indicator age bracket of 14 to 16, for example.

In general, the ages between 15 and 35 suffer relatively low mortality overall, limiting the backcasting error even if those with more education have a moderate survival advantage. Perhaps the largest selection effect in some settings will be differential migration, as the more educated are more likely to migrate abroad.

The model backcasts completion rates at a_3 (the top of the indicator age bracket) for individuals aged up to 20 years above the nominal age bracket. The error is assumed to generally increase with increasing distance in age.

Nuisance factor

When the average completion of the age group $[a_{l+3}, a_{l+5}]$ is calculated based on empirical observations, this corresponds to the population-weighted average

$$\sum_{n=l+3}^{l+5} \frac{p_n}{p_{[l+3,l+5]}} a_n$$

The variation in p_n does not provide useful information regarding school completion but represents a nuisance factor. 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 $[a_{l+3}, a_{l+5}]$ will significantly exceed true differences in birth cohort size in all but the largest surveys and extreme fertility settings.[iv]

Moreover, differences in single year cohort size for the *backcast* values will further be distorted by random variation in mortality and migration. For backcast values based on different surveys, nothing at all can be learnt about the relative size of single year cohorts.

Finally, recall that one of the purposes of the model is to *project* completion rates to the current year. 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.

Accordingly, we calculate, backcast, estimate, and project synthetic completion rates that represent the *unweighted* averages

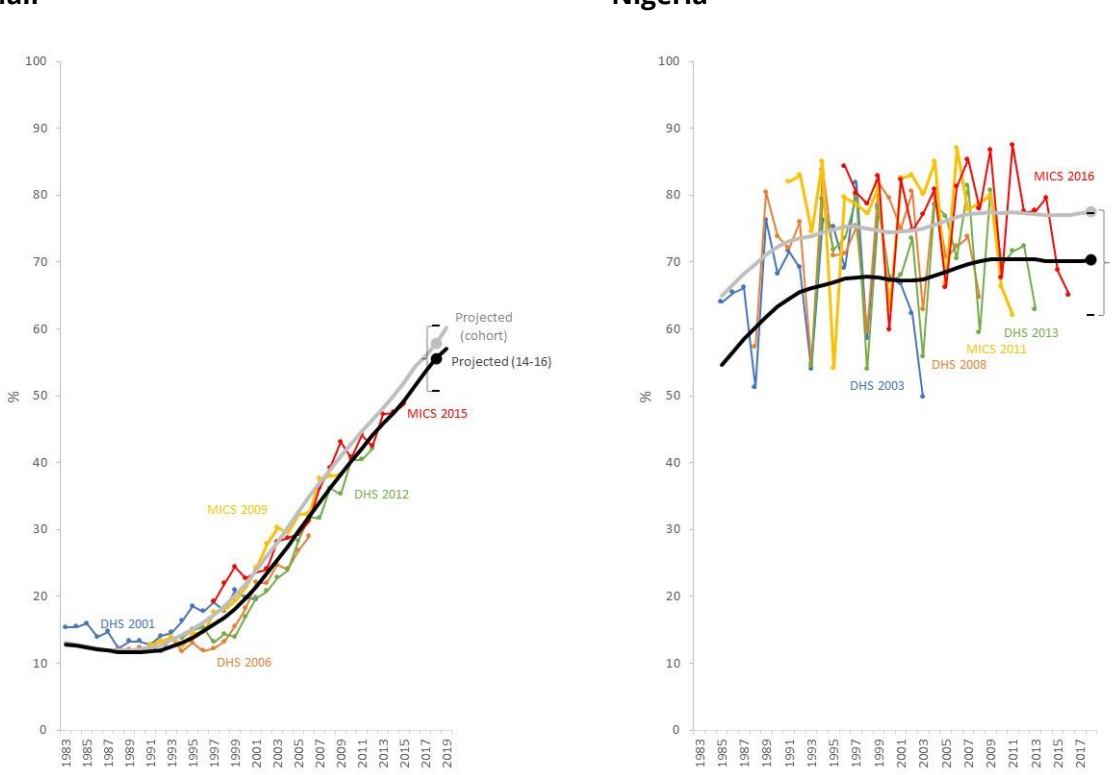
$$\sum_{n=l+3}^{l+5} \frac{1}{3} a_n$$

While this approach is preferable in terms of estimation and understanding the true trend in education system performance for the reasons explained above, this means that our values for the completion rate estimate based on a given survey may differ marginally from those customarily reported for the survey year. However, the age-specific $\widehat{C}_{a,c,y}$ can be combined with exogenous projections and historical imputations of population estimates by single years of age to replicate and/or project $\widehat{CR}_{c,y}$.

Results

The effect of late completion is evident in a comparison of two countries in **Figure 3**. In Mali, the observed values show little evidence of late completion, while the primary completion rate has rapidly expanded since the mid-1990s. In Nigeria, by contrast, there is wide variation in the observed values of the completion rate both between and within household surveys. There is also a discrepancy between the fitted completion rate for the target age group 14- to 16-year olds (black line) and the completion rate of the cohort, which includes people who complete primary education at a later age; the standard completion rate indicator may be underestimating the cohort completion rate by 7 percentage points. Regardless, Nigeria is characterized by stagnation in the primary completion rate since the late 1990s.

Figure 3. Observed, fitted and projected primary completion rate, Mali and Nigeria, 1983–2018



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

The model will be compared to three alternative, simpler specifications in terms of predictive performance with respect to the latest survey observations. This will take the form of a 'leave one out' validation. Specifically, all observations based on the latest survey (including backcast values) are omitted from the estimation of the models, and predicted values for these values are obtained.

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 $\mathrm{CR}^\star$, 'latest' is the only alternative model specified in terms of the $\mathrm{CR}^\star$. Instead, the other benchmark models are, just like the main model, based on partially backcast age-specific estimates of C. The first of these is the 'flat' model M^f. This fits an intercept-only least squares model to the κ for a given country. The second is the 'simple' model M^s that models the κ for a given country as a linear function of an intercept and a slope over time.

However, since the presence of late completion is obvious in many countries, it is clear that current observations of completion at young ages cannot be directly compared to back-projected values. Accordingly, since a simple baseline model will, by definition, not model late completion explicitly, the

most reasonable approach appears to be to perform validation *only* on the back-projected 'observations'.

Conclusion and next steps

Global model-based estimates of development indicators that are based on survey data should not be mistaken for real. But they can be invaluable in (i) supporting a better understanding of trends, helping make sense of disparate sources, and (ii) using available information efficiently to make short-term projection to the current year.

The next steps in the analysis involve:

- a review of the methodology and its robustness (see section 3) with a view towards adopting this methodology for reporting on the indicator for SDG 4 purposes
- a discussion of the relative merits of projecting the completion rate for the originally intended group (e.g. 14- to 16-year olds in primary education) and the cohort completion rate, which tends to be higher
- the analysis of survey sources in addition to the 80 countries that draw on DHS and MICS data, in order to produce regional and global averages

Annex: Core model

The probit of the true age-specific completion rates at the top end of the relevant age interval, $C_{a_l+5,c,y}$, are modeled as following a random walk with drift over time, with auto correlated shocks. Formally, the de-trended first differences follow a stationary AR(1) process. With Φ the cumulative density function (CDF) of a standard normal distribution, $\kappa_{a,c,y} = \Phi^{-1}(C_{a,c,y})$, and

$$\Delta\kappa_{a_l+5,c,y} = \kappa_{a_l+5,c,y} - \kappa_{a_l+5,c,y-1} = \gamma_c + \delta_{c,y}$$

where $\delta_{c,y} \sim \mathcal{N}(\rho\,\delta_{c,y-1}, \tau)$ is normally distributed with variance τ.

As additional robustness checks, the sensitivity of the results to the values of fixed hyper parameters, as well as to the length of the backcasting window, will be investigated.

---

[i] Accordingly, a parsimonious model for late completion is preferred. Specifically, completion between ages a_{l+3} and a_{l+5} is assumed to potentially be lowered by a country-specific constant age slope. Formally:

$$\kappa_{a_l+4,c,y} = \kappa_{a_l+5,c,y} - \lambda_c$$

$$\kappa_{a_l+3,c,y} = \kappa_{a_l+5,c,y} - 2\lambda_c$$

[ii] The estimation is based 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. Some DHS and MICS survey reports provide sampling error estimates for selected key indicators. However, none do so for school completion rates as defined here. All sampling errors have therefore 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. Specifically, the sampling variance of any given observed completion rate $\widehat{C}_{a,y,c,s}$ in year y at age a in country y from survey s is estimated as (omitting indices for clarity):

$$\widehat{Var}\,(\hat{\kappa}) = \frac{1}{k(1-k)}\sum_{i=1}^{k}(\hat{\kappa}_i - \hat{\kappa})^2$$

where

$$\hat{\kappa}_i = kc - (k-1)\,\hat{\kappa}_{(i)}.$$

Here, $\hat{\kappa}$ is calculated on the full sample, $\hat{\kappa}_{(n)}$ is calculated on the sample with the n^{th} cluster excluded, and k is the total number of clusters.

[iii] In developing countries, respondent age is often misreported, leading to an overrepresentation of ages that are multiples of five. Age misreporting tends to be negatively associated with school completion: reported primary school completion was 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. Reconstructed observations that represent a reported 'round' age group at the time of survey are coded with an indicator variable. Observations where this indicator equals 1 are subject to an additional term in the model equation that models the potential distortion due to age misreporting. This takes the form of a threshold model: a country-specific Bernoulli outcome with prior B(0.5) determines whether age misreporting occurs; if it does, the magnitude of the distortion is modelled as a Gamma(0.5,1) distribution. 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 modelled explicitly as affecting specific ages, but is allowed to be implicitly absorbed in the overall country intercept.

[iv] In 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 $\sqrt{(20{,}000 \cdot 0.02 \cdot 0.98)}/20{,}000 \approx 0.001$, or 5% in relation to the true value of 0.02.