WP 17.1
24 October 2013

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint Eurostat/UNECE Work Session on Demographic Projections
organised in cooperation with Istat
(29-31 October 2013, Rome, Italy)

Item 17 – Population projections by age and sex and level of education

Estimating transition age schedules for long-term projections of global educational attainment

Bilal Barakat, Wittgenstein Centre for Demography and Global Human Capital

Estimating transition age schedules for long-term projections of global educational attainment

Bilal Barakat
Wittgenstein Centre for Demography and Global Human Capital
bilal.barakat@oeaw.ac.at

Prepared for the Joint Eurostat/UNECE Work Session on Demographic Projections, 29 - 31 October 2013, Rome, Italy

DRAFT — PLEASE DO NOT CITE

1 Introduction

The present exercise forms part of a larger project of generating a new set of global population projections disaggregated by educational attainment. These projections make use of education-specific fertility, mortality, and migration rates. In other words, the education serves as an input into a multi-dimensional cohort-component population model. At present, in the absence of an explicit model of intergenerational transmission, the dynamic time trend in educational composition of subsequent cohorts is not generated endogenously, but is projected independently. The general approach underlying these education projections is presented here. A particular focus rests on the specific sub-task of determining the *timing* of attainment progression to be assumed for projection, given that it cannot be estimated directly from the cross-sectional baseline data. While the question of timing has limited impact on the projected adult populations in the long run, it has a significant impact on the projections in the short term and for younger age groups.

2 Long-term projections of global educational attainment

Our overall approach consists of two parts: fitting a model of educational development trajectories to the empirical development of attainment over the

course of recent decades; and projecting attainment by extending these trajectories into the future.

The specifics of both parts are interdependent. Estimating the model is logically prior, since it is meaningful without attempting a projection but the converse is not true, and the feasible characteristics of the model constrain the form of the projection. Examples of such constraints are discussed below. Conversely, the intended primary purpose of the model as a projection input determines the requirements. On the one hand, the model must be of a kind that can be extrapolated in a meaningful and consistent way; on the other hand, it need only capture temporally robust associations, not causal explanations. The model is fit to reconstructed educational attainment data for 178 countries for the period 1970–2010.

2.1 Fitting past expansion

The first question is which measure(s) to model directly, and which to derive. In the present context, educational growth could conceivably be modelled by projecting the number of people attaining each level, or the shares of different attainment groups among each cohort, or transitions between levels. There are arguments in favour of each choice. Modelling absolute numbers accounts for some of the absolute constraints on capacity expansion in the education sector, and might suggest a view that educational development is supply-driven. Transition rates between attainment levels arise naturally from the common approach to modelling within-school flows in terms of entry, promotion, repetition, and drop-out rates. We choose to focus on attainment shares among (five year) cohorts as the measure to be modelled directly. The distinct advantages are three-fold: firstly, in a world of declining fertility, supply-side constraints are expected to lose importance relative to social demand, especially at levels below tertiary, and normalising the education model by population size makes it possible to specify it independently from the overall population growth model and its fertility assumptions. This breaks a potentially troublesome feedback loop in the projection method. Secondly, in contrast to transitions, attainment is directly a property of persons, and represents the outcome of the education model required for inclusion in the overall population projection (this also serves as an argument against examining enrolments, which are indeed a person characteristic, but not a stable one). Moreover, shares at different levels can be interpreted independently, whereas transitions are cumulative and their implications for attainment can only be understood as an ensemble.

Having settled on attainment shares as the principal measure, the question arises whose attainment to model directly, or more specifically, which age group to focus on. It would be possible, of course, to separately model the attainment at different ages, but the possibility of contradictory results makes

it an unattractive approach. Using a single reference age group, however, calls for a balancing act: younger ages reflect more closely the recent developments in the education system, while at older ages there is a greater chance of accurately capturing the maximal lifetime attainment. Here, we use the attainment share at ages 30–34 as the benchmark, in a compromise between these two concerns.

An additional question concerns the levels to model explicitly. Separate models for different attainment levels risk creating inconsistent results, where the share of those with upper secondary or higher attainment is projected to be greater than the share with lower secondary or higher attainment, for example. This risk is greater the closer the levels in question are to each other. Instead of attempting a complex joint model that imposes ordering constraints, we estimate independent models for the sufficiently separated levels of completed primary, upper secondary, and tertiary education, which results in projections without inconsistent crossovers. We then interpolate the intermediary levels of lower secondary and incomplete primary education, with no education calculated as the residual. Specifically, the share of these intermediary levels among those with completed primary, but less than completed upper secondary schooling (in the case of lower secondary), and among those with less than completed primary schooling (in the case of incomplete primary), respectively, is held constant at the most recent value observed.

The model is designed to capture the intrinsic dynamics of the education sector, rather than the effect of external predictors (such as economic growth), that would then, in turn, need to be projected, raising additional questions of data reliability and endogeneity.

The model assumed here effectively relates the growth over time in the share of 30–34 year olds having reached a given attainment level or higher to the current level of said share. The relationship is such that growth is slowest for both very low and very high shares and fastest at middle levels, resulting in a sigmoid, trajectory over time. This is not an axiomatic assumption, but borne out by empirical observation. To illustrate, the next figure displays the observed country paths at the original scale, arranged around the hypothetical global average trajectory. While there are of course numerous examples of irregular and/or stalled expansions, the graph for South Korea shows that individual countries too may follow the archetypical sigmoidal pattern.

2.2 Model specification

In the present case, this shape is parameterised as an (inverse) probit curve. The probit curvature was found during exploratory investigations to match the empirical patterns more closely than a logistic specification, for instance. Concretely, the observed data on the highest attainment are transformed into reversely-cumulated attainment shares such as "percentage with upper sec-

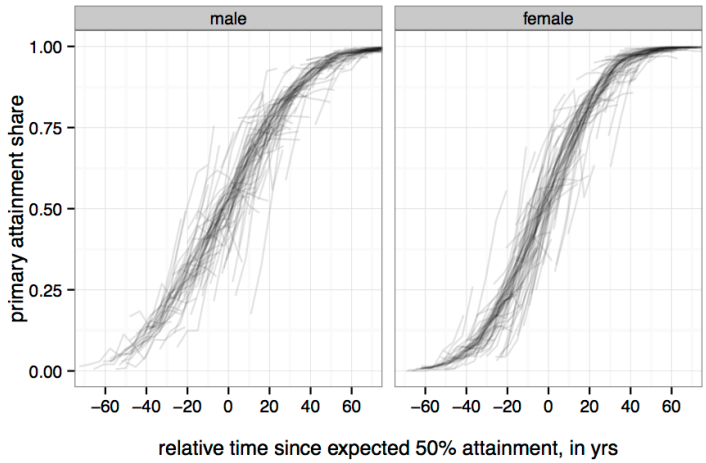


Figure 1: Global pattern of expanding primary attainment in the age group 30–34

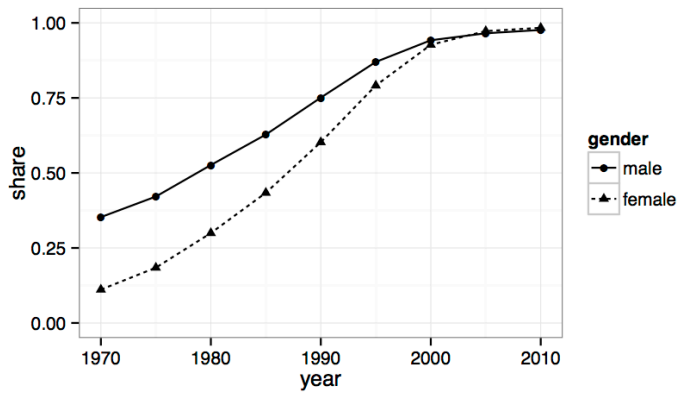


Figure 2: Individual country example. Here: lower secondary attainment at age 30–34 in South Korea, a frontrunner for expansion.

ondary or higher" and correspondingly for other attainment levels. These figures in the interval [0,1] are probit-transformed into unbounded numbers. An exact inverse-probit sigmoid-curve would be perfectly linearised as a result, turned in other words into a straight line. Accordingly, a linear fit to the transformed data corresponds to an inverse-probit sigmoid-curve fitted to the original data. The linear predictor is taken to consist of an overall global component g, as well as region-specific and country-specific elements (r and c), all time-invariant, as well as random residuals epsilon at the level of country-year dyads. Formally:

$$\text{participation}_{it} = \Phi(x_{it})$$

$$x_{it} = x_{i(t-1)} + g + r_i + c_i + \epsilon_{it}$$

Because we are not interested in lateral shifts in time, the data can be centred so that participation is 50 percent at time t_0 for estimation purposes. In order to obtain distributional results to aid scenario creation, the above specification is estimated within a Bayesian framework. The priors for all parameters are normally distributed with zero means and half-Cauchy priors for the variances. The outcomes of the estimation are posterior probability distributions for the overall global, regional, and country-specific rates of educational growth.

For the post-secondary education level, the attainment share is re-scaled so that complete saturation corresponds to 90% of a cohort attaining post-secondary education. This reflects the fact that, unlike lower levels of schooling, universalising post-secondary education is nowhere a policy target, and, on the contrary, the current frontrunners such as Singapore are debating whether to actively limit post-secondary expansion.

The next figure shows the empirical pattern of attainment growth by level. The dashed line is a locally weighted scatterplot smoother (LOESS). The solid line displays the predicted growth based on only the global average term g. In other words, this represents the global average inverse-probit shaped expansion path. As is evident, even this simple, purely endogenous, model does a credible job of approximating the pattern in the data.

2.3 From modelling the past to projecting the future: additional considerations

Even given a perfect model of the past, deriving projections for the future calls for the consideration of issues that go beyond a mechanical extrapolation of the estimated trajectories. Two issues stand out: questions of convergence, and how to generate alternative scenarios. These are not orthogonal, of course, since convergence may be interpreted as a specific scenario. Nevertheless, it is useful to discuss these issues separately.

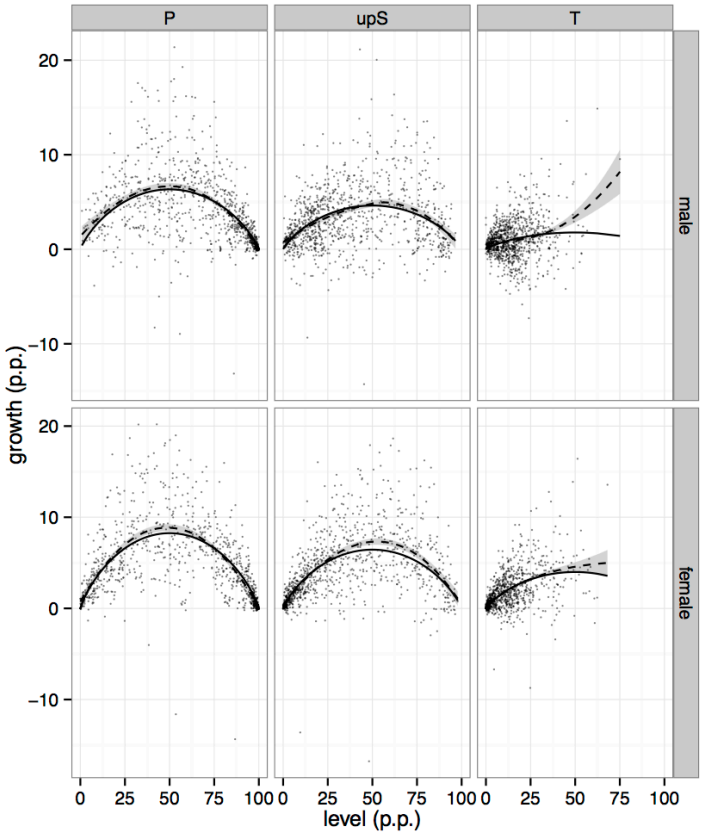


Figure 3: Observed (dots), smoothed (dashed), and predicted (solid) 5-year attainment growth (in percentage points) as a function of attainment share already achieved.

Convergence

Due to the specification of the model, where countries' expansion parameters are estimated as coming from a shared statistical distribution (by gender and education level), the basic country estimates and, by implication, their projected trajectories, are not independent. Some "shrinkage" towards the overall mean occurs, reflecting the fact that a country with an exceptionally fast/slow historical expansion path may be assumed to have experienced a particularly fast/slow incidental spell in addition to having a fast/slow intrinsic momentum.

However, the estimation model does not assume that countries become more similar over time. This is, nonetheless, something that may be assumed as a projection assumption. Here, it is the rates of change in probit-transformed attainment that undergo convergence, not the attainment levels directly. In

other words, even if complete convergence were assumed, countries would still differ in attainment, and–due to the non-linear sigmoid expansion model–would still expand at different rates on the original scale of attainment shares, because even with identical rates on the transformed scale, countries at middle levels of attainment would increase their attainment share more rapidly than countries close to saturation.

The case for assuming a weak convergence over a very long time horizon is different from the case for assuming relatively rapid complete convergence (within a few decades). The first is a "regularisation" of the projection rather than a reflection of a strong assumption of mechanisms leading to convergence. In the absence of any convergence, countries that have undergone a recent decline in educational attainment would be projected to undergo an educational collapse if the decline were extrapolated without a corrective. Assuming a slight level of convergence to the global median ensures that such countries' trajectories are merely stagnant. Accordingly, a slow convergence in rates, namely complete convergence by 2100, is assumed for all scenarios. By contrast, the central scenario is one where globalisation trends in educational development result in a convergence in rates by 2060.

A separate issue is that of convergence of educational attainment of males and females within a given country. The initial estimates of the trajectories are independent for each gender. Indeed, historical patterns show that large gender differences in attainment can occur and remain for decades at all education levels. Gender difference in primary school attendance in countries with low educational development in sub-Saharan Africa, South Asia, and elsewhere are systematically highlighted by international development agencies. But large gaps at the post-compulsory stage can also be observed in industrialised countries. Nevertheless, for projection purposes, it is appropriate to include some degree of linkage between female and male attainment levels in a given country to avoid a situation where gaps are projected that fall outside the range of precedent. Here, male and female attainment, in terms of the share attaining or exceeding each level of education, are projected to converge to the gender-averaged attainment by the end of the projection horizon in 2060.

Scenario definition

Basing the projection on the median estimates of the country slopes corresponds to a "business as usual" setting. One of the aims of the exercise is to investigate the possible consequences of more rapid, or on the contrary more laggard, educational growth.

In the central scenario, the attainment profile of future cohorts is based on the median parameter estimates of the model estimated above. In that sense, while not interpretable as the "most likely" scenario in a probabilistic sense, it can be interpreted as the scenario that reality is equally likely to exceed or fall short of.

In policy terms, this may be interpreted as "business as usual". This does not, however, imply a static perspective. On the contrary, even if the pace of educational expansion statistically depends on the level already attained in an endogenous fashion, it is clear that this expansion, though statistically unsurprising, nevertheless has to be actively produced by the actors involved. It is therefore a scenario of sustained effort. At the same time, being based on the average performance of the most recent decades, including at the country-specific level, this scenario does factor in the inevitable setbacks and mismatches between ambitious policy targets and actual change "on the ground". One of the aims of the exercise is to investigate the possible consequences of more rapid, or on the contrary more laggard, educational growth.Numerous distinct ways of defining scenarios of more rapid growth based on the above model exist:

- chose a higher quantile, such as the 75th percentile, from the estimated distribution of country slopes, instead of the median, to estimate the future development. Everyone "ups their game". This is indeed what is displayed in the example trajectories shown below.
- Adjust the trajectories to ensure every country meets some defined thresholds, such as 70 percent upper secondary attainment by 2040, in analogy with current targets at lower levels. Think "MDG 2".
- Countries with a below-average pace of growth converge to the global median. "The fast stay fast, the slow become average".

These schemes could easily be combined by determining for each country the pace implied by each of the schemes, and applying the highest, or the middle one of them (or even the lowest, for a "bearish bullish" scenario). Each of these can be inverted for a pessimistic scenario. The first and third more readily than the second, though. For a "slow growth" equivalent to the second scheme above, it seems more natural to assume lower asymptotic saturation levels instead.
As it were, for the larger project a different, entirely non-model-based approach was chosen. In particular, the high scenario was based on all countries following the expansion path of the past frontrunners such as South Korea. This is therefore not to be interpreted as an "optimistic expectation", but as an upper bound. Conversely, the low scenario is based on constant attainment shares.

3 Timing of attainment progression

Our aim is to obtain projected attainment levels for all age groups for the period 2010-2060. The above model in the first instance provides these only for the age group 30–34. Given the simplifying assumption that educational attainment is mostly complete by age 30-34, changes in the attainment shares

at higher ages result from differential mortality and migration behaviour between different education groups, rather than attainment transitions. Accordingly, the shares at higher ages derive from the general multi-state population projection and require no further input from the education projection itself. Attainment projections for age groups younger than 30 depend on the timing of attainment at different levels. This creates the challenge of specifying attainment age schedules that are coherent from a cohort perspective, and consistent with both the baseline data for younger age groups and the projected time series of completed attainment (Fig. 4).

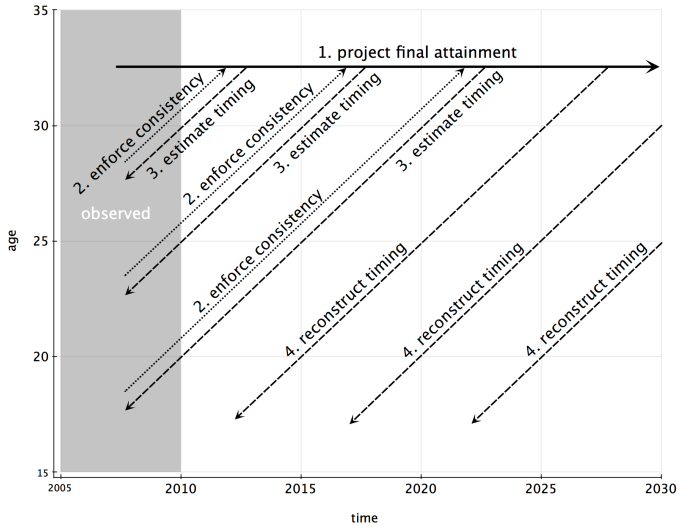


Figure 4: Short-term reconciliation between projected attainment at age 30-34 and observed attainment at younger ages

3.1 Enforcing logical consistency between empirical baseline and raw projection

To elaborate on the last point: Because it is time series of attainment at age 30-34 that is explicitly modelled and projected into the future, there is potential for disagreement with the observations of attainment at younger age groups. In principle, it is possible that the cohort 25-29 in the base year already has higher levels of attainment than the simple projection implies for the age group 30-34 five years later.

This affects only the first three projection steps, as the individuals who in the base year are located in the age groups 25-29, 20-24, and 15-19 have observed attainment shares that might be inconsistent with their projected attainment at age 30-34 based on the simple projection. Beginning with the cohort that

is aged 10-14 in the base year, there are no observations that could be at odds with the projected attainment at later ages.
A two-step process ensures consistency along these two directions. The first is along cohort lines. Since the highest level of education attained is non-decreasing with age, this is a logical requirement. It is achieved by adjusting the projection so that the share having achieved a given education level in a cohort is at all ages at least as high as the share observed in the base year.
The second step ensures that the application of the first step does not result in fluctuations over time in attainment at a given age. While this is not a logical necessity since educational stagnation and even decline across cohorts are certainly possible in reality, this is done in the interest of coherence. Fluctuations around the central trend are possible, but the entire model is set up to project the long-term central trend, not short-term fluctuations.
Accordingly, the maximum is taken of the values resulting from a) the simple projection of attainment at age 30–34 over time, starting in the base year, and b) the attainment of younger age groups observed in the base year, brought forward along cohort lines.
In effect, the local maximisation approach to the first step corresponds to the assumption that, as far as possible, inconsistencies should be reconciled by adjusting the assumed timing of attainment, rather than the final level. Consider, for example, a situation in which the observed attainment at age 20–24 is higher than the projected attainment at this age, but lower than the projected final attainment at age 30–34. Our approach reconciles this by adjusting the timing of attainment so that sufficiently many attainment transitions are assumed to have occurred earlier than initially assumed, but without changing the overall number of transitions to higher levels. In principle, it would be equally possible to fix the timing schedule, in other words the assumed relationship between attainment at age 20–24 and final attainment at age 30–34, but to adjust the assumed final level so that the back-projected attainment at the younger age no longer falls below the observed level. The choice between the two approaches (and against a mixture of them) was dictated by the principle that the least certain parts of the model should be the first to be adjusted to accommodate contradictory information. In the present case, the timing schedules are reasonable a priori assumptions, while by contrast the projected final attainment levels at age 30–34 also incorporate modelling assumptions, and hence are also data-driven.

3.2 Age schedules for attainment progression

In an ideal situation, with time series data for attainment by age, it would be possible to estimate a set of age schedules empirically. Unfortunately, the estimation is woefully underdetermined in the present situation where the baseline data consist of a single cross-section. Inevitably, some structural assumptions

must be made that constrain the estimation. The following section discusses the specification, followed by a presentation of our approach to combining *a priori* assumptions, the projections for the age group 30–34, and the empirical data on younger age groups.

3.2.1 Specification: Markovian rates versus "teleological" conditional schedules

The baseline model might be considered to consist in the specification of a set of 6x6 transition matrices indexed by age. If these are assumed to fully capture the probability structure of the transition behaviour, the underlying assumption is that the rates of educational progression depend only on current age and education level already achieved. In particular, this excludes path dependence, in other words that late achievers differ from early achievers in their subsequent behaviour. Discounting structural zeroes arising from the monotonicity (by definition) of the highest level attained and sum-to-unity constraints, this gives rise to 15 independent transition rates for each age group, or 45 parameters overall.

In a single-year model, additional zero constraints may be implied if it assumed that "leapfrogging" several levels within a single time period is not feasible. In a five-year model, such an assumption is more difficult to justify. Since we are dealing not only with five-year time steps, but also with aggregated five-year age groups, it is possible that an individual completes several levels even if the sum of their durations exceeds five years, since the older members of the age group may in the first period be observed just before completion of the next level. For example, an 18 or 19-year-old member of the 15–19 age group may be observed in the final year of, but just before completion of, upper secondary school in time period t, and have already completed both secondary school and a four-year tertiary degree by time t+5. Similar arguments apply to larger jumps. In particular, individuals aged 15–19 without completed primary schooling who do progress normally gain their primary school or even lower secondary school certificate in some kind of accelerated programme, rather than by attending a 4–6 year course of schooling. As a result, the only "leapfrogging" we can exclude with some confidence is the maximal case of a transition from no schooling to completed tertiary education within a single five-year time interval. Taking this into account reduces the degrees of freedom merely to 42 instead of 45.

Constraining this fully flexible model is necessary both for pragmatic reasons of estimability, and in order to take into account our domain knowledge. With regard to the former, note that in the present case the observed baseline data only contains a single observation of the attainment distribution for each age group. In other words, 15 data values in total, once residual categories are accounted for. For now, let us disregard the fact, discussed in detail below, that these represent cross-sectional period data and therefore do not

necessarily allow the estimation of transition rates along the life course if these are not static. Regardless of this issue it is clear that the fully flexible model fails to be determined by the available data.

It therefore becomes necessary to encode additional assumptions in the model. In principle, it would be possible to do so without violating the Markov property, by specifying functional relationships between the transition entries within each age-specific matrix. However, this approach is not followed here. For one, it is not obvious that we possess relevant background knowledge that would allow us to state, for example, that at a given age the transition rate from completed lower secondary to completed upper secondary be identical or proportional to the transition rate from completed primary to completed lower secondary. In addition, we positively do not expect the Markov property to hold in reality. What does correspond to our domain knowledge is that students are heterogeneous, and that this heterogeneity affects, among other things, the timing of their educational progress. For example, we expect a high school graduate who completed each school phase "on time" to have a different probability of eventually completing tertiary than a student who repeated several grades in primary school, even if both are observed with complete upper secondary schooling at age 20–24. In this spirit, it would be possible in principle to encode these expectations in a set of *a priori* age-specific transition schedules for indexed by a relatively small number of different "types". The inference problem then reduces to estimating the mixture of types in the population.

However, in the context of the larger projection exercise, such an approach carries several disadvantages. Firstly, since we are interested in a dynamic situation and are projecting the final educational attainment distribution to change between cohorts, we must also find a way to specify the time evolution of the transition schedules. If a set of type-specific schedules are used, this amounts to projecting how the distribution of types changes in the future. Without a substantive interpretation of what these types represent, this is hardly a meaningful undertaking. What we do have, independently of the transition schedules by age, is a projection of final attainment by age 30–34. Secondly, regardless of how we arrive at the estimated transitions, one step necessary to complete the "back-projection" of educational attainment is to distribute the individuals whose attainment later in life is known over attainment levels in younger years. Given a set of "forward" transition rates, this requires the calculation of conditional "backward" schedules. In other words, given that an individual has attainment X at age 30–34, what is his or her probability of having been at level Y at age 25–29? And so on at younger years. Thirdly, if there are N types, this approaches requires up to $42N$ parameters to be fixed *a priori*, to leave $N - 1$ degrees of freedom for fitting the model to empirical data.

3.2.2 A set of model conditional schedules

Together these two considerations suggest an alternative approach, based on specifying a set of *a priori* "backward" schedules directly, indexed by final attainment reached. The intuition behind this approach is captured in the corresponding elicitation questions. These follow the scheme of: "Given that some individuals are known to have completed upper secondary school, but not tertiary, by age 30–34, what share of them can be expected to have completed primary/lower secondary/upper secondary by age 15–19/20–24/25–29 respectively?" This involves specifying six matrices with six education levels by three age groups each. Taking sum-to-unity constraints and "hard" structural zeroes implied by monotonic attainment progression into account, this calls for 45 parameters to be specified *a priori*, to leave a model with $6-1=5$ degrees of freedom to fit to the data. This not only represents a fraction of the $6 \times 42 = 252$ parameters that would be necessary to specify a "forward" schedule model with 5 degrees of freedom (see above), but most of the 45 backward parameters are much more constrained by subject considerations. While the types, in the absence of a substantive interpretation for them, do not by themselves impose any clear constraints, and in addition, transition rates are in principle independent of each other, the backward distributions are subject not only to structural zeroes due to monotonic progression, but also to numerous inequality constraints. Moreover, conditional on final attainment, an assumed structure of "normal" durations of different education levels readily translates into timing schedules conditional on final attainment. For example, given a final attainment of tertiary education, a 6+3+3+4 model of primary, lower secondary, upper secondary, and tertiary duration, together with a starting age of 6, implies that primary was completed around 12, lower secondary around 15, upper secondary around 18, and tertiary around 22 years of age. The actual schedule assumed is naturally a marginal deviation from this standard. Arguably, at most a dozen parameters are actually debatable, the remainder being naturally either zero or 100 percent, even if they are not *logically* constrained to be. The aggregation into five-year age groups complicates matters: even in the simplest case where everyone completes the upper secondary level at the nominally standard age of 18, for example, the age group 15–19 would include both individuals between the ages 15 and 17 who will complete the level but only in the future, and those aged 18 and 19 who have already graduated. The specification of attainment timing in terms of five-year age groups should take this into account.

The implications of the duration structure for transition *rates* are rather more complex, because the transition rate of 16-year-olds to completed lower secondary depends not only on the normal durations, but on the share who actually progress. Put differently, the backwards model cleanly separates timing of educational attainment (given by schedules conditional on final attainment) from overall level changes (shifts between schedules). In sum, the backward

specification corresponds more with the framing of domain knowledge, incorporates changes over time more easily, is more aligned with the tasks of the overall projection exercise, and is more parsimonious in a setting where there is no microsimulation of individuals, only of evolving aggregate shares. A set of proposed model schedules is included in the appendix.

3.2.3 Schedule estimation

The transition schedules are estimated for each country by taking the short-term projections for the age group 30-34 as given, and comparing them to the observed attainment at ages 25-29, 20-24, and 15-19 in the baseline year. Specifically, if 30 percent are projected to have completed post-secondary attainment or higher by age 30-34 in 2015, and 20 percent are observed to do so among 25-29 year olds in 2010, then the implied schedule is that, of those completing post-secondary, two-thirds do so between the ages of 25-29 and 30-34. A similar comparison is made for the age group 30-34 in 2020 and the age group 20-24 in 2010, and for the age group 30-34 in 2025 and the age group 15-19 in 2010.

These schedules are not guaranteed to be consistent from a cohort perspective. In a second step, schedules are determined through optimisation techniques that satisfy logical consistency constraints and are closest in absolute difference (over the entries) to the raw schedules found above. More precisely, a weighted average distance is minimised between the raw country-specific schedules and a standard set of model schedules described, with weights of 0.7 on the country-specific schedule and 0.3 on the model schedules. This regularisation step avoids some undesirable transition behaviour that can occur in the raw schedules in countries whose baseline population attainment at younger ages is highly irregular due to rapid educational change, declining attainment, or possibly migration. The assumed schedules are designed to correspond to nominal graduation ages based on school entry at age six, with a stylised school system of 6+3+3 for primary, lower, and upper secondary schooling, and three years for a first post-secondary degree, while taking into account the age spread within five-year age groups. At the same time, it must be recognised that at the post-secondary level, there is no consensus that a "regular" education career requires higher education studies to immediately follow the completion of upper secondary schooling. Indeed, "late" post-secondary attainment, even above the age of 35 is substantial in some countries, including the USA.

The sensitivity of the population projections with respect to the assumptions concerning attainment timing is expected to be greater in the case of fertility than for mortality. Overall mortality is generally lowest at the young adult ages at which the assumptions apply. The greatest impact is expected to be on fertility and migration. However, this impact is moderated by the fact that absolute educational fertility differentials are smallest in the countries

with high levels of post-secondary attainment and largest in those countries where the share of post-secondary graduates, and therefore the share of those whose attainment timing is most difficult to estimate, is relatively low. In general, the specification of assumptions in terms of age schedules conditional on ultimate attainment means that, in contrast to the alternative of specifying transition rates, specification errors do not cumulate.

3.3 Results/Examples

The above procedure provides partly empirical, partly model-based estimates of the age schedule for attainment transitions in younger years, on a per-country basis. As described above, the balance between the two components is determined by a mixture parameter. The following figure illustrates the effect of changing the parameter, and motivates the decision to use a value of 0.7 for the definitive estimation. Displayed is the attainment distribution for the age groups 15–19, 20–24, 25–29, and 30–34 for parameter values in the range 0.5–0.75, both five and twenty years into the projection, for Burundi. It is clear that the baseline data (left column, independent of the parameter) are highly irregular and do not by themselves yield cohort transitions that are logically consistent, or that are plausible after merely enforcing consistency. It is also evident how in the medium term, strong rectangularisation (top row) serves to iron out the irregularities altogether. While in this particular case, a strong shift occurs between the values 0.50 and 0.55, the location of this regime switch differs between countries. The value of 0.7 was picked after inspection as providing a sensible compromise between maintaining country differences, but removing artefacts such as large shares of extremely late attainment.
Regarding the final output of the overall projection exercise, the following figures display the results for attainment shares (across both genders) at age 30–34 for Nepal. It can be seen that the model provides projections that smoothly extend the past trajectory, while also taking into account the ceiling effect in a natural way. Moreover, in this particular case, the past trajectory suggests that Nepal is on an expansion path that is somewhat more rapid than average, but not dramatically so, as evidenced by the fact that the median country-specific projection is consistently higher than the one based on the global trend alone, but that the latter falls within the high-low range of country-specific trajectories. This observation is an important factor in assessing the plausibility of the trajectories. For example, taken out of context, it may seem excessively optimistic to project that more than 80 percent of Nepalese 30–34-year-olds will have completed lower secondary school in 2050, but this is put into perspective by noting that even a merely average trajectory by international comparison would already yield a number of around 75 percent.

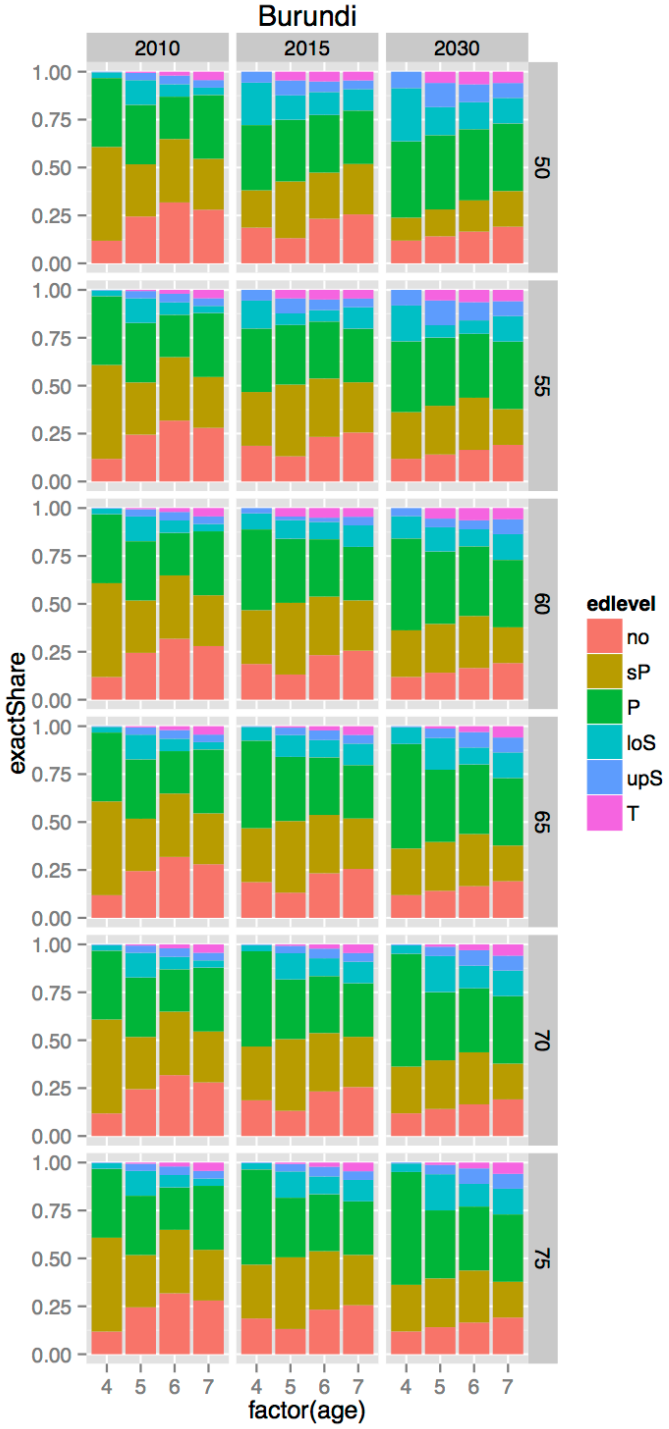
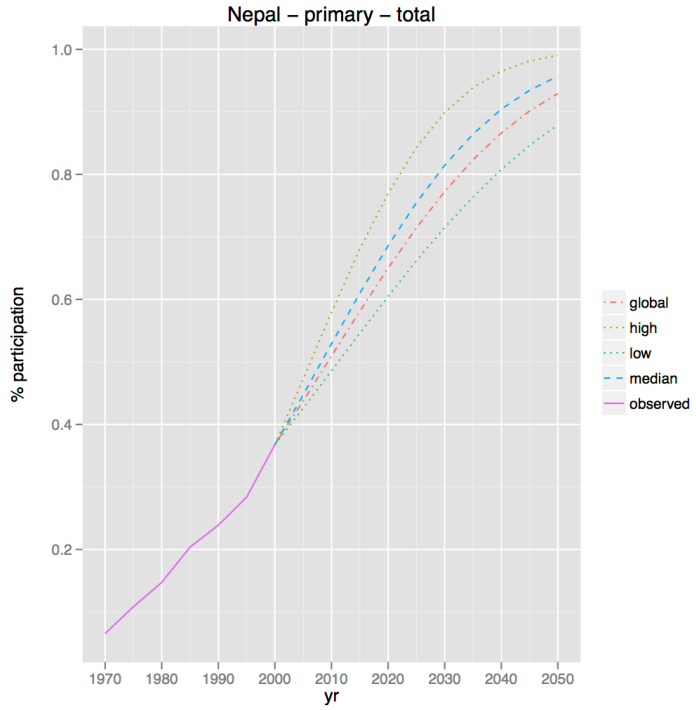
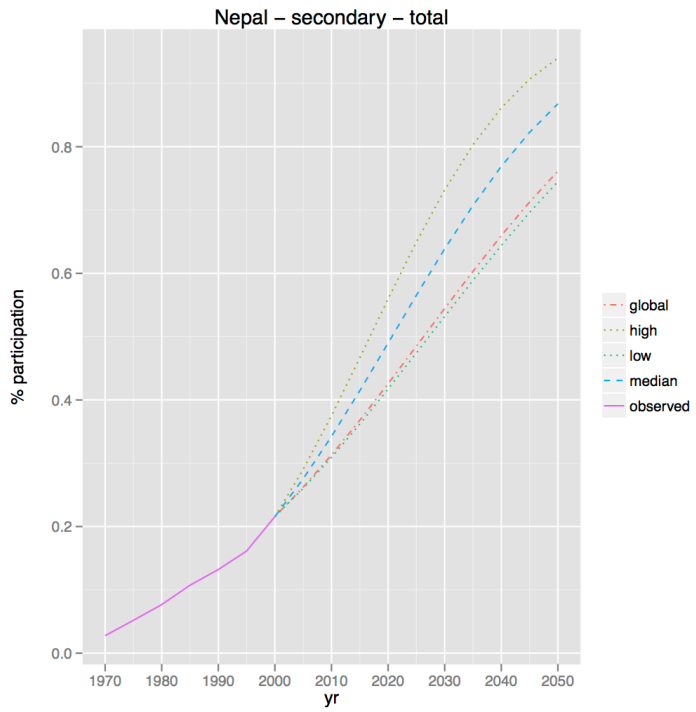


Figure 5: Projected attainment diffusion in Burundi for different amounts of normalisation.


Nepal – primary – total
% participation
1.0
0.8
0.6
0.4
0.2
1970
1980
1990
2000
2010
2020
2030
2040
2050
yr
global
high
low
median
observed


Nepal – secondary – total
% participation
0.8
0.6
0.4
0.2
0.0
1970
1980
1990
2000
2010
2020
2030
2040
2050
yr
global
high
low
median
observed

4 Limitations and possible extensions

Since the baseline data are in many cases derived from censuses, sampling variation is only a marginal consideration when deviations of the observed data from model schedules are considered. A fully Bayesian implementation where the mixture weights derive from likelihoods would therefore mostly serve to account for model uncertainty. However, it is clear that such a specification would have to be performed at the individual country level, since in some countries the duration of different school levels is *known* to be different from the "standard model" assumed above, whereas in others it is almost exactly correct. It is therefore not an alternative that could easily be applied wholesale across countries, if the priors really are to encode real information.

Given a longer time-series of past observations, it would be possible in principle to attempt to *estimate* whether convergence between countries or genders is occurring, rather than relegating the question to one of making a scenario assumption for projection. In a similar vein, a random walk specification is arguably more defensible than the above assumption of independent period deviations around a central trajectory. After all, a lull in educational expansion may indeed create constraints for future expansion, in terms of the availability of suitable teachers but also through dynamics of inter-generational transmission.

Nevertheless, even the simple approach presented here that relies on some assumptions that are somewhat ad hoc, even if not entirely arbitrary, demonstrates that it is feasible to exploit period data, together with projection methods, to generate a) country-specific transition schedules for the timing of educational attainment, and b) projections of the final attainment status, that are both consistent from a cohort perspective and reasonable from an educational perspective.

5 Appendix

5.0.1 A set of model age schedules for educational attainment

level	15–19	20–24	25–29
tertiary	0	70	100
upper secondary	40	30	0
lower secondary	60	0	0
complete primary	0	0	0
incomplete primary	0	0	0
none	0	0	0

Table 1: conditional on tertiary at age 30–34

level	15–19	20–24	25–29
upper secondary	40	100	100
lower secondary	60	0	0
complete primary	0	0	0
incomplete primary	0	0	0
none	0	0	0

Table 2: conditional on upper secondary at age 30–34

level	15–19	20–24	25–29
lower secondary	100	100	100
complete primary	0	0	0
incomplete primary	0	0	0
none	0	0	0

Table 3: conditional on lower secondary at age 30–34

level	15–19	20–24	25–29
complete primary	100	100	100
incomplete primary	0	0	0
none	0	0	0

Table 4: conditional on complete primary at age 30–34

level	15–19	20–24	25–29
incomplete primary	100	100	100
none	0	0	0

Table 5: conditional on incomplete primary at age 30–34

level	15–19	20–24	25–29
incomplete primary	0	0	0
none	100	100	100

Table 6: conditional on no schooling at age 30–34