Vienna Yearbook of Population Research 2007, pp. 193-235

Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000

Wolfgang Lutz, Anne Goujon, Samir K.C. and Warren Sanderson*

Abstract

Using demographic multi-state methods for back projecting the populations of 120 countries by age, sex and level of educational attainment from 2000 to 1970 (covering 93 percent of the 2000 world population), this paper presents an ambitious effort to reconstruct human capital data which are essential for empirically studying the aggregate level returns to education. Unlike earlier reconstruction efforts, this new dataset jointly produced at the International Institute for Applied Systems Analysis (IIASA) and the Vienna Institute of Demography (VID) gives the full educational attainment distributions for four categories (no education, primary, secondary and tertiary education) by five-year age groups and with definitions that are strictly comparable across time. Based on empirical distributions of educational attainment by age and sex for the year 2000, the method moves backward along cohort lines while explicitly considering the fact that men and women with different education have different levels of mortality. The resulting dataset will allow new estimates on the impact of age-specific human capital growth on economic growth and first results show—unlike earlier studies—a consistently positive effect.

* Wolfgang Lutz (author for correspondence), World Population Program, International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Austria and Vienna Institute of Demography, Austrian Academy of Sciences, Vienna, Austria. Email: lutz@iiasa.ac.at

Anne Goujon, Vienna Institute of Demography, Austrian Academy of Sciences, Vienna, Austria and World Population Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.

Samir K.C., World Population Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.

Warren Sanderson, Departments of Economics and History, State University of New York at Stony Brook, USA and World Population Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.

DOI: 10.1553/populationyearbook2007s193

1 Introduction

This paper is the first report of an ambitious, multiphase project whose aims include the production of a new national level dataset on educational attainment by age and sex for as many countries in the world as possible over the period 1970-2000, the analysis of these new data, the making of projections of educational attainment by age and sex for those countries up to 2050, and the assessment of the likely effects of future changes in educational structure. The project is a joint effort of the World Population Program at the International Institute for Applied Systems Analysis (IIASA) and the Vienna Institute of Demography (VID). The first version of educational attainment reconstructions is now complete. We call it Version 1 because it is a complete set of data that was produced following the rules specified in this paper and that went through a first round of validation of results. In the future there will be more detailed validations and possible country-specific adjustments of our assumptions that ultimately will result in a Version 2. But Version 1—as described in this paper—is now ready to serve as input for a first round of analyses. In this paper we describe the methods used for reconstructing the educational attainment distributions for 120 countries using the methods of multi-state demographic modelling. The data itself will soon be available with full details on the Internet.

For many years economists interested in the determinants of economic growth have been puzzled by the fact that indicators of the human capital of a population sometimes do and sometimes do not show significant positive coefficients in cross-sectional and time series regressions of economic growth as would be expected according to economic theory. This lack of consistent empirical evidence on the macro-level returns to education is in stark contrast to the strong evidence on the individual level where it is well established that more education on average leads to higher income. This unsatisfactory situation has lead to the suspicion that the problem may not lie with the theory or the models used but rather with the aggregate level education data themselves. If the puzzle of inconsistent micro and macro effects of education on the economy can be solved by using more accurate, consistent and detailed education data, this would be a major contribution to economic growth research. Indeed, first applications of standard economic growth regressions to these new data showed consistently significant positive coefficients for human capital and in this sense performed much better than previous human capital datasets (Crespo and Lutz 2007).

This reconstruction exercise focuses strictly on levels of educational attainment, which are measures of the quantity and formal level of education received. Educational quality also has an important effect on human capital. Standard measures of skills acquired such as the PISA or PIRLS school

performance databases[1] or the International Adult Literacy Survey (IALS) for adults are based on actual testing of samples of the population and show strong variation between countries that could explain other differentials associated with education. However, such datasets based on direct testing of skills are so far only available for a small number of countries (mostly member countries of the Organisation for Economic Co-operation and Development (OECD)) but efforts are under way (e.g., by the United Nations Educational, Scientific and Cultural Organization's (UNESCO) Institute for Statistics (UIS)) to collect such information for a larger number of countries. In the future we plan to incorporate educational quality and skills assessed on the basis of testing into our measures for countries where data are available, but this will be done in a later phase of the project.

Following this introduction, this paper has six sections: Section 2 introduces the basic idea of demographic back projections and discusses earlier applications. Section 3 discusses the existing data sources and earlier reconstruction efforts, which were mostly based on economic perpetual inventory methods. Section 4 contains the main body of the paper, describing our method. It begins with a concise summary of the different steps involved and then discusses at some length the key dimensions of the method: the raw data and their adjustment, the assumptions about mortality differentials and migration, our ways of dealing with the open-ended age group and with the age at progressing to higher attainment categories, and finally the assumptions needed to convert the reconstructed attainment distributions into mean years of schooling. Section 5 gives a brief discussion of selected results and Section 6 presents some sensitivity analyses. The concluding section will give a short outlook of what kinds of studies are now made possible with these new data, and what we plan to do as a next step.

2 The approach of demographic back projections

Comprehensive assessments of the returns to investments in formal education at the aggregate (national) level as well as other studies of the impacts of human capital require empirical information about the educational status of the adult population over some period of time for a large number of countries. This information needs to be consistent in terms of the definition of educational categories across countries and over time. Since the effects of educational attainment can also be expected to differ by age (e.g., one might expect that the education of 25-34 year olds should be more important for economic growth than that of persons beyond retirement age) as well as by sex, having full age details

[1] PISA (Programme for International Student Assessment) measured the performance levels of pupils aged 15 in reading, mathematical and scientific literacy in 2000, 2003 and 2006. PIRLS (Progress in International Reading Literacy Study) was conducted in 2001 and 2006 to measure the reading and comprehension skills of pupils in the fourth year of primary education.

for men and women can be considered a great asset for a comprehensive analysis. In addition, only the explicit consideration of distinct levels of educational attainment allows for the analysis of the relative importance of primary versus secondary or tertiary education (and different mixes of the three) which should be key to the development of relevant education policy plans at national and international levels. Such consistent information by age, sex and level of education has not been available so far for a large set of countries, including both industrialized and developing countries and over several decades of time, although some partial efforts at reconstructing levels of educational attainment have been developed at a more aggregated level.

In this section we briefly describe the general approach taken in producing this new human capital dataset. Unlike earlier reconstruction efforts that mostly used economic capital accumulation models, this joint effort by IIASA's World Population Program and the VID is based on demographic multi-state methods that allow vital rates in different educational categories to differ. Starting with only one empirical dataset for each country for the year 2000, we go back in time and reconstruct earlier distributions by level of education along cohort lines. Since the overall size and age distribution for each country and point in time is given by the population estimates of the United Nations (UN) Population Division, the task of this reconstruction effort essentially was to estimate the proportions with different educational attainment for each given five-year age group of men and women for the period 2000 back to 1970.

The concept of projecting populations backward in time is not new. Applications have mostly been in historical demography for reconstructing population size and structure for early periods for which no such information was otherwise available. Wrigley and Schofield (1982) developed a specific back-projection method to provide new demographic estimates for England for the period 1541-1871. A method of 'inverse projection' had also been developed by Lee (1978) to estimate demographic structures in the past. In a later paper, Lee (1985) performed a critical appraisal of the Wrigley and Schofield 'back-projection' technique and modified his own 'inverse-projection' technique in order to be able to perform the same task done by Wrigley and Schofield and compare the results.

One of the tasks in the Wrigley and Schofield work was to estimate the population sizes and age distributions in the past from a recorded series of births and deaths and a terminal age distribution, say at time *t*. The method first estimated the number of deaths occurring in the oldest closed age group[2] during time *t-5 to t* using data on respective cohort sizes of the oldest closed age group at times *t* together with some assumptions. The number of deaths is then used to find

[2] An open age group typically covers a broader age interval than a closed age group and its end value on one side is not specified, e.g., 65+ is an open age group as opposed to 65-69 which is a closed age group.

the model life table generating the number, which is then used to reverse-survive all age groups except the oldest one.

A problem arises when the number of reverse-survived aged 0-4 does not match the number of births in the previous years. Wrigley and Schofield attributed the difference to migration. These migrants need to be distributed over the cohort's life span and hence affect the estimates of the age distribution at previous steps, and consequently the estimates of previous mortality levels.

In addition to this problem there are certain assumptions to be made to obtain consistent mortality levels and numbers of death in the oldest closed age group. The method requires iterations to arrive at a consistent estimate. In general, the key issue with back projection outlined by Lee "… is how to estimate the number of people in the oldest closed age group each time one moves back a step in time…" (Lee 1985: 236). These methods are in principle quite similar to our method, the difference being that our task is not to estimate the age structure (which is given by the UN) but rather the educational distribution for each given age group that requires the consideration of education-specific mortality and migration levels.

In a different context, the method of demographic back projection has been used widely to estimate HIV incidence from AIDS incidences data (see De Angelis, Gilks, and Day 1998, cited in Law et al. 2001) and to estimate the number of dependent heroin users from the observed numbers of opioid deaths and new entrants to methadone treatment (Law et al. 2001). For these applications the task is, in general, to estimate the number of people in an initial state given the information about the number of people in the final state, and make assumptions about the rates of progression to the final state.

The basic idea of back projection in the context of reconstructing the educational distribution is rather simple: Assuming that the educational attainment of a person remains invariant after a certain age, we can derive, e.g., the proportion of women without any formal education aged 50-54 in 1995 directly from the proportion of women without formal education aged 55-59 in 2000. Assuming that this proportion is constant along cohort lines, it directly gives us the proportion of women without education aged 25-29 in 1970. In a similar manner, the proportions for each educational category and each age group of men and women can simply be moved to the next younger five-year age group as one move back in time in five-year steps. It is important to see that these are not arbitrary assumptions, but truisms under certain conditions. In the above example, the proportions of women without schooling aged 25-29 in 1970 and 55-59 in 2000 must be identical if nobody moves to the category with primary education after the age of 25 and if mortality and migration do not differ by levels of education. This follows directly from the fact that the size of a birth cohort as it ages over time can only change through mortality and migration. In reality we know, however, that mortality tends to strongly vary with the level of education in every country of the world and that migration can do so as well in specific cases.

That is why we—unlike earlier reconstruction efforts—will make special adjustments for these differentials as will be discussed in the following sections.

It is worth noting that we do not have to worry about the level of fertility. Typically, fertility assumptions are a key concern in population projections, in particular with respect to education, as fertility tends to be sensitive to a woman's level of education and is typically much higher for uneducated women than for highly educated women. In a forward projection, the size of a population increases through births and in-migration and decreases through deaths and out-migration. Conversely, in a backward projection, the population increases along cohort lines by accounting for mortality and migration. The level of fertility can be indirectly inferred from the size of the youngest age group but does not enter as a component of change when going backward in time. However, if we have reliable independent information about the number of births in the past (e.g., from birth registration) we could assess the accuracy of our mortality and migration assumptions in our back projections by comparing the reconstructed age group 0-4 with the child-mortality adjusted number of children aged 0-4 according to the birth statistics. But for this specific back-projection exercise, even such considerations are irrelevant because we only project the population down to a minimum age of 15 (because we focus on educational attainment) and also because the age and sex structure (without the education detail) is not reconstructed but directly taken from the UN estimates.

Formally our model can be summarized as follows: Starting with $t = 2000$ as the jump-off year for our back projection for which we have a full distribution of the population by age (five-year age groups), sex and level of education (four categories), when there are no transitions between education levels, we go back in time in five-year intervals calculating the same full distribution for year t-5 according to

$$N(age-5, educ, t-5, sex) = \frac{N(age, educ, t, sex)}{SurvivalRatio(age-5, educ, t-5, sex)} \quad (1)$$

where

N(.) refers to the number of people in the group defined by (.),
age refers to the five-year age group starting with age *a* (e.g., *a*=20 refers to the age group 20-24),
educ refers to the educational attainment category (see definition below),
t refers to calendar year *t* and *t-5* to five years earlier,
sex refers to the gender of individuals,
SurvivalRatio(.), refers to the proportion of people surviving for five years in the country (i.e., combining mortality and migration) in each age-, sex- and education-specific group over the period *t*-5 to *t*.

The aim of the back projection is to obtain a dataset with the population distributed by five-year age groups, starting at age 15 and up to the highest age group 65+, by sex, and by four levels of educational attainment over a period of 30 years from 2000 (base year) back to 1970 in five-year intervals.

The four educational attainment states (ISCED refers to the International Classification of Education) are defined as:

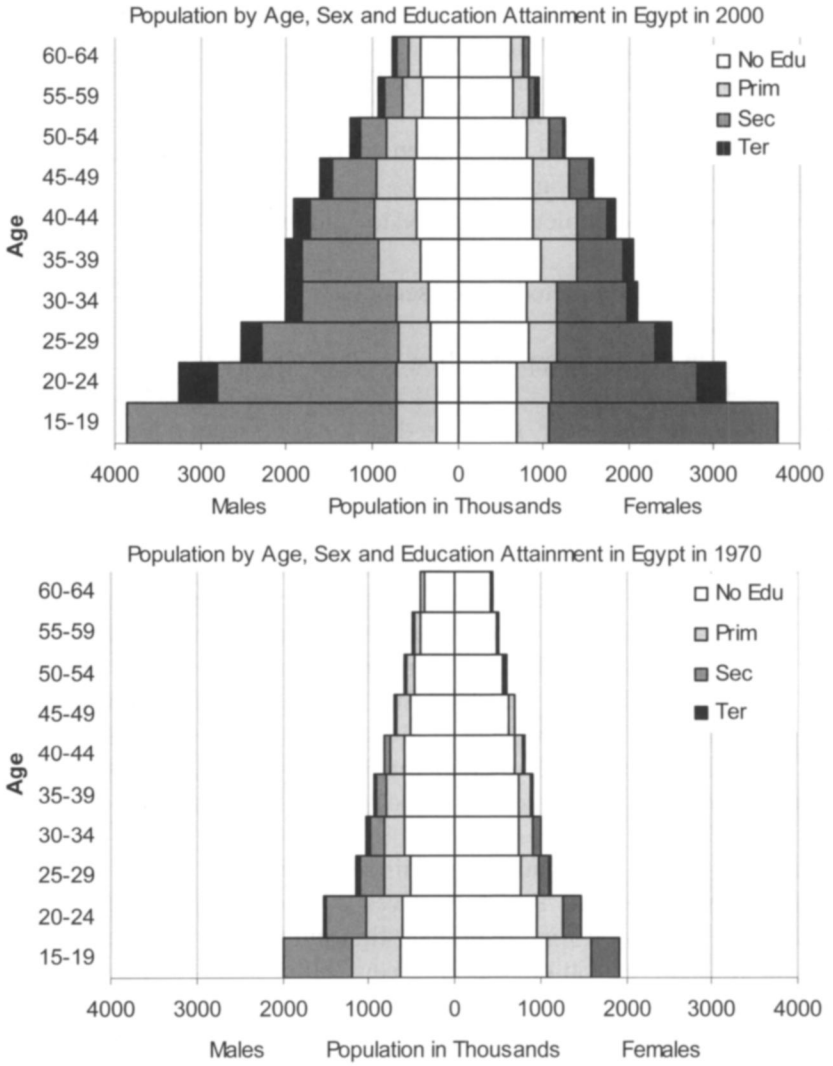
- No education: those who have never been to school and have received no formal education (No Education);
- Primary: those with uncompleted primary to uncompleted lower secondary (ISCED 1);
- Secondary: those with completed lower secondary to uncompleted first level of tertiary (ISCED 2,3 and 4);
- Tertiary: those with at least completed first level of tertiary (ISCED 5 and 6).

We chose 2000 as the base year, since the data for 2005 were not available for a vast majority of countries. Our method completely depends on the educational input in the base year. This makes the baseline education-related data very important, since no other inputs on education are introduced during the back projection, unlike earlier reconstruction efforts that often used school enrolment rates. This makes the model very dependent on the quality of the baseline data for the year 2000, but has also the great advantage that the educational attainment categories by definition cannot change over time, which has been the main stumbling block for using the empirical UNESCO data and earlier reconstruction efforts. Since our empirical baseline data is always standardized in terms of the age and sex distribution to exactly match the UN data, only the part of the empirical information that refers to the education distributions is of critical importance.

To illustrate the kind of information that this reconstruction method generates for 120 countries in the world, Figure 1 gives an example in terms of age pyramids by level of education for Egypt. The first pyramid shows the structure by age, sex and level of education for the year 2000, which is the empirical baseline information used for the reconstruction. The second pyramid gives the reconstructed structure for the year 1970, resulting from our method. The pyramid for 2000 shows that educational attainment for the younger cohorts in Egypt has been improving recently. While more than half of the women above age 35 had no formal education, in the age group 15-19 more than 80 percent of the women have been to school. The education profile in 1970 resembles that of the population above age 45 in 2000, which is the basis for its reconstruction.

Before we describe this method and the critical assumptions involved in more detail, we will have a look at the existing empirical data in this field and discuss previous efforts to reconstruct the missing information.

Figure 1:
Age pyramids by level of education for Egypt for 2000 and 1970



3 Existing data and previous reconstruction efforts

When trying to collect empirical international data on educational attainment by age and sex over time, it is at first surprising to see how little consistent time series data on levels of educational attainment exist. This is not only the case for developing countries but also for developed countries with established statistical offices and routinely organized censuses. Two main problems hinder the

availability of a database that is consistent over time. The first is the definition of the categories for which data has been collected. Here the problem lies not so much with changes in the education systems themselves but rather with changes in the definition of categories used for collecting statistical information about the level of education. This is especially true of censuses carried out in the 1960s and 1970s. Secondly, although most countries around the world have an education system organized along the same general structure of primary, secondary (junior and high/vocational and general), tertiary (vocational and general), a comparison across countries becomes difficult when considering the differences in the length of the different cycles: Is a four-year primary education in Kuwait equivalent to a seven-year primary education in Mozambique? Cycle length hides another problem that cannot be addressed simply by examining levels of educational attainment. The problem lies in the curriculum and the quality of education affecting the comparability of students in terms of their skills at the end of a cycle. Some particular surveys have noticed substantial differences. For instance, the completion of primary education in some African countries does not necessarily entail even the achievement of full literacy skills. However, and as mentioned in the introduction, measuring levels of educational attainment represent an important first step in the development of a concise and consistent database.

Because of the high importance of consistent international time series on the human capital of the adult population, several efforts have been made to construct such series using whatever exists in terms of available empirical data. The problem is that the official data from censuses such as those collected by the UN Statistical Office and UNESCO are only fragmentary and scattered over time and countries. In addition, these data suffer from various changes in definitions of educational categories over time and across countries, which make them inappropriate for consistent time series analysis. Despite intensive efforts by UNESCO in terms of harmonizing the data, data collection is still a national responsibility with censuses carried out at different points in time, countries have their own statistical traditions reflecting the specifics of their education systems and an imperfect process of communicating census results to the relevant international bodies, which in some cases are raw and in others adjusted census data and often they do not contain the necessary age detail.

UNESCO (and more recently the newly founded UIS) has traditionally provided the main source of data on levels of educational attainment. Together with the UN Statistical Office, census data on educational attainment has been collected since the 1960s. Those data were generally published in the annual UNESCO yearbooks for aggregate age groups (mostly 15+ or 25+) since the late 1960s, showing more age detail in special issues (e.g., 1978, 1988, 1995 and 1997).[3] The data in the UNESCO databases suffers from all the problems present

[3] Age- and sex-specific levels of educational attainment were also published in the United Nations Demographic Yearbooks (Special topic: Population census statistics).

in the original data as mentioned above. Another difficulty is the fact that, for the sake of consistency, national data are further classified according to UNESCO's predefined categories for all countries and the allocation of the census data to the UNESCO categories may have caused some of the observed inconsistency problems. This is complicated by the fact that UNESCO has incorporated changes in their definition of categories according to the changes made by the international standard classification ISCED. An important change was implemented in recent years and is particularly problematic for the reconstruction of consistent time series. Since around 2000, the data on the highest educational attainment levels are based on completed levels of education, with the categories being, no schooling, incomplete primary, completed primary (ISCED 1), completed lower secondary (ISCED 2), completed upper secondary (ISCED 3) or completed post-secondary, non-tertiary (ISCED 4) and tertiary completed (ISCED 5 or 6). Older data until the end of the 1990s were collected in terms of participation in the levels from secondary upward and contained no information on completion. Those categories were no schooling, first level (non-complete/completed), entered second level (S-1 and S-2) and post secondary.

Because of the inconsistent and fragmentary nature of the purely empirical dataset collected from national census information, several attempts have been made in the past to estimate complete, comprehensive and consistent datasets for large numbers of countries. Table 1 compares the three most important such datasets to our newly reconstructed one in terms of selected key features, such as age detail, educational categories, number of countries, time coverage, etc. The first and most often used dataset was developed by Barro and Lee (1993, 1996, 2001) who complement the existing attainment data with the somewhat more consistent time series of national school enrolment data at different levels using perpetual inventory methods which help transform accumulated education flows (enrolment) into human capital stocks. This resulted in a widely-used dataset that gives the proportion of the population by highest level attained and mean years of schooling of the entire adult population (by sex but without age details) for 142 economies, of which 107 have complete information at five-year time intervals from 1960 to 2000. The main drawback of the Barro and Lee methodology is that the authors used existing real data and interpolated gaps based on enrolment rates, making the data very sensitive to inconsistencies in the educational categories used, as mentioned above. Similar independent efforts have been made by Kyriacou (1991), Lau et al. (1991), Nehru et al. (1995), De la Fuente and Doménech (2006) and by Cohen and Soto (2007), which in many cases result in quite different estimates of mean years of schooling, with most of the estimates being significantly higher than Barro and Lee. A recent summary of available educational datasets can be found in Cohen et al. (2007) and Bloom (2006). None of the listed reconstruction efforts give the desirable age detail cross-classified with the distribution over different educational attainment categories. They also disregard in their calculations the well established fact that people with higher

Table 1:
Comparison of the characteristics of selected major reconstruction efforts of levels of educational attainment for larger numbers of countries

	Barro and Lee	**De la Fuente and Doménech**	**Cohen and Soto**	**IIASA/VID**
Age groups	Two large age groups: 15+ and 25+	One large age group: 25+[a]	One large age group: 15-64[a]	5-year age groups: 15-19; 20-24; …65+
Sex	Male/female/total	Total	Total	Male/female/total
Education indicators	Proportions by highest level attained + MYS[b]	Proportions by highest level attained + MYS[b]	Only MYS[c]	Proportions by highest level attained + MYS[b]
Period covered	1950-2000[c] (5-year steps)	1960-1995 (5-year steps)	1960-2000 (10-year steps)	1970-2000 (5-year steps)
Specific educational categories used	7 categories: No schooling; first level (total/complete); second level (total/complete); post secondary (total/complete)	6 categories: Illiterates; primary schooling; lower and upper secondary; first and second cycle of higher education	Not mentioned	4 categories: No schooling; primary; secondary; tertiary
Coverage in terms of countries	107 countries (and 142 countries with partial data)	21 OECD countries	95 countries	120 countries
Empirical data source used	Censuses and enrolment series	National sources (censuses, surveys)	OECD, censuses, Mitchell Series	Censuses, DHS[d], LFS[e] for year 2000
Methodology used	Perpetual inventory method, interpolation	Proceeding backward from 1990 or 1995 by backward and forward interpolation, or rely on miscellaneous information	Extrapolate backward – assumption of constant proportions assumed. Net School Intake Rate used in case of no census data	Reconstruct 5-year age groups along cohort lines from 2000 backwards considering mortality/ migration differentials

Notes: [a] Age groups are used during calculation but not presented in the resulting database.
[c] Data for 2000 result from projections.
[b] MYS stands for Mean Years of Schooling.
[d] DHS refers to Demographic and Health Surveys.
[e] LFS refers to Labour Force Surveys.

education have lower mortality rates, which can have quite significant effects on the educational composition of the older adult population, as will be demonstrated in the sensitivity analysis section below. One common disadvantage of all these exercises (with the notable exception of Barro and Lee and De la Fuente and Doménech for OECD countries) is that the main indicator used is mean years of schooling (MYS). This indicator is used in most of the numerous economic growth regression models that have been produced over the past years. The calculation of MYS, which requires many assumptions, will be discussed in Section 4.6. However, it hides the potentially important effect of educational attainment distributions.

While all these previous reconstruction attempts have made important contributions to the discussion, only our new reconstruction is fully comprehensive in the sense that it provides full age detail (five-year age groups) cross-classified with the educational attainment distribution for a large number of developing and industrialized countries. Moreover, due to the specific approach chosen, our method is insensitive to the problem of changing educational classifications over time because we only use the classification given for the empirical data in 2000 and project those backward in time. However, this makes our data sensitive to the quality of the 2000 data. The general assumption here is that data collection has improved over time and that most information collected in recent years is more reliable than that collected earlier. Also careful checks were implemented to choose the most reliable data when several sources were available (for instance, a census and an LFS). Of course, the reconstruction does not come without certain assumptions, which we will discuss in detail in the remaining parts of this paper. But at this point it is also important to stress that the data as included in our database has been validated, i.e. for every country the data has been compared to existing historical sources of data and in some cases changed accordingly as will be described later. In this sense our new data also reflects those other independent sources of historical information.

A detailed country-level comparison of our results with those of the most important other datasets has recently been carried out, but goes beyond the scope of this more methodological paper. The findings from the comparisons will be published in a forthcoming paper. In this context it should be mentioned that in terms of overall average levels of education, our data are closer to those of Cohen and Soto and De la Fuente and Doménech than to Barro and Lee, which on average show significantly lower levels than the majority of other datasets.

4 Our method

To give the reader an overview of the method before going into some of the details, Section 4.1 presents a formal summary of the procedure we used. The raw data and the adjustments made to them are discussed in Section 4.2. In Section 4.3 we present the assumptions about differential mortality and migration by education that are used in the reconstructions. Section 4.4 deals with the procedures for dealing with open age intervals. In Section 4.5 we discuss the assumptions that we used with respect to age-specific educational category progression rates. Section 4.6 presents the assumptions that were used in computing mean years of education, which is a derived indicator from our reconstruction results and which is produced primarily to facilitate comparison with other studies and as a service to users who prefer to capture human capital by a single indicator.

4.1 Summary of procedure used

Box 1 summarizes the key steps taken in producing our reconstruction results for all 120 countries:

Step 1: Find reliable empirical information on the proportions of population by levels of educational attainment for men and women for five-year age groups for the base year (around 2000).

Step 2: Adjust the educational categories, if necessary, to make them comparable across countries.

Step 3: Apply the empirical proportions to the age structure as given by the United Nations Population Division (UN 2005) for the corresponding country for the year 2000.

Step 4: Obtain the *period life expectancy* **at age 15** for all men and women from the UN general model life table as used for the corresponding country for the period 1995-2000, i.e., the five-year period preceding *t* (Source: UN 2005).

Step 5: Calculate the corresponding **education**-specific *period life expectancy* at age 15 by using education differentials in life expectancy as described in Section 4.3.

Step 6: Obtain survival ratios for all five-year age groups above age 15 corresponding to each education-sex-specific period life expectancy at age 15 (using the UN general model life table).

Step 7: If there is no empirical information for the closed age interval 65-69 but only for the open interval 65+, the information for 65-69 must be estimated according to the procedure described in Section 4.4.

Step 8: Calculate the number of people *N(age,educ,sex,1995)* by age (*age* going from 15-19 to 60-64), sex and education living five years earlier (in 1995) by using Equation. (1) above.

Step 9: Adjust for the transitions to secondary and tertiary education that happen after the age of 15 as described in Section 4.5.

Step 10: Convert the number of people by age and education calculated for 1995 (*t*-5) into age- and sex-specific proportions and apply to the UN (2005) estimates of population structure for this year in order to assure full consistency (including adjustments for migration).

.... Go back to Step 4 and repeat the procedure until the year 1970 is reached.

4.2 Raw data and their adjustments

Our goal was to include as many countries as possible in our analysis with the selection criterion being the availability of reliable baseline data. So far, we have been able to obtain such information for 120 countries, but we will aim to expand the coverage as new information becomes available for additional countries. What is considered to be satisfactory baseline information must, of course, be subject to some degree of judgment.

For each country, our reconstruction methodology requires an initial distribution of the population by sex and age (by five-year age groups starting at age 15 to at least the age groups 60-64 and 65+) in 2000. We searched for such data and were able to collect the data for 120 countries. Our main sources were national censuses mostly from UNESCO, but also directly from national statistical agencies, Demographic and Health Surveys (DHS), and Labour Force Surveys (LFS). But even these data were not always in the form we needed. The main irregularities stem from data referring to years slightly different from the year 2000, data that have only 10-year age groups, data where the last age group was lower than 65+, and data with differing educational attainment categories.

We dealt with the problem that not all empirical data pertain exactly to the year 2000 by introducing a two-year tolerance limit for the time to which the information refers, i.e., accepting data referring to the years 1998-2002. If we only had data for the years 2003-2005 or 1995-1997, we applied backward or forward projections along the lines described here to bring all countries to the common starting line of 2000.

In more detail, we obtained our empirical data for the starting year from the following sources: the database of the UIS (35 countries), DHSs (33 countries), Eurostat (16 countries) and LFSs (eight countries). These data were complemented by census data provided by national statistical offices (NSO, 27 countries). For China we used Microdata (a sample from the year 2000 census). The specific sources of data for each country as well as the adjustment procedures that were used to iron out some of the irregularities are documented in all necessary detail in the database itself (they are also accessible as an appendix to this paper at http://www.iiasa.ac.at/Research/POP/edu07/index.html).

One of the main problems that had to be solved before we could estimate consistent starting data was the inconsistency between educational attainment categories used in the DHS and our categories based on the new ISCED standard reflecting completed levels. Since there was a sufficiently large number of countries with information from both DHS and censuses following ISCED, we established a relationship between the classification schemes as described in Table 2. A set of adjustment factors was estimated based on the regression of the 10 countries for which recent UNESCO and DHS were available (Armenia, Brazil, Côte d'Ivoire, Guatemala, Jordan, Namibia, Peru, South Africa, Tanzania and Turkey), which would translate the DHS categories into our categories. The

DHS proportion for "no education" was kept the same because this is the only identical category. Other proportions were multiplied by the adjustment factors and further adjusted in a second step to bring the sum of all proportions (without changing the no education proportions) to unity. Those final adjustment factors are listed in Table 2.

Table 2:
Differences between IIASA/VID categories based on ISCED and DHS categories, plus the adjustment factors used

Category/Data Source	IIASA/VID	DHS	Adjustment Factor[a]
No education	E1	1	1
Some primary	E2	2	1.15
Completed primary			
Some lower secondary	E3	3	1.24
Completed lower secondary			
Some higher secondary			
Completed higher secondary			
Some tertiary education		4	0.60
Completed tertiary education	E4		

[a] The adjustment factor was multiplied to the DHS data across all age groups for both males and females.

Using this procedure, we estimate the starting populations by age, sex and four levels of attainment and visually display the results using multi-state age pyramids as shown, for example, in Figure 1 for Egypt for the year 2000. Such a visual representation gives the main features of the distribution at a glance. Figure 1 shows that adult women are significantly less educated than men and that for both men and women, the educational attainment is much better for the younger cohorts. The shape of the pyramid also shows the sizes of the cohorts indicating that for Egypt, the younger adult cohorts are not only better educated, but also much more numerous than the older ones. This is the case in many developing countries that have experienced improved education over the past decades. It is also clearly visible for India (see Table 3 and Appendix A).This fact by itself will lead to significant improvement in the educational composition of the adult population, even if school enrolment rates do not increase in the future, simply because the more educated, more numerous cohorts will move up the age pyramid over time and replace the less educated, smaller ones. Multi-state forecasts by level of education for India clearly demonstrate this phenomenon (see Lutz and Scherbov 2004).

Table 3:
India around 2000 (data from the 2001 census). Proportions of the population with four educational attainment categories for men and women by age

	Males				Females			
Age	No education	Primary	Secondary	Tertiary	No education	Primary	Secondary	Tertiary
15-19	0.17	0.27	0.56	0.00	0.29	0.24	0.46	0.00
20-24	0.19	0.22	0.50	0.08	0.40	0.21	0.33	0.07
25-29	0.24	0.23	0.41	0.12	0.48	0.21	0.24	0.07
30-34	0.28	0.24	0.37	0.11	0.54	0.21	0.20	0.06
35-39	0.32	0.26	0.32	0.09	0.58	0.21	0.17	0.04
40-44	0.34	0.27	0.30	0.09	0.61	0.20	0.15	0.04
45-49	0.34	0.27	0.30	0.09	0.64	0.19	0.13	0.03
50-54	0.38	0.26	0.27	0.08	0.69	0.18	0.10	0.03
55-59	0.39	0.29	0.25	0.07	0.75	0.16	0.07	0.02
60-64	0.49	0.28	0.18	0.05	0.81	0.13	0.05	0.01
65-69	0.49	0.31	0.17	0.04	0.81	0.14	0.04	0.01
70-74	0.54	0.29	0.13	0.03	0.84	0.12	0.03	0.01
75-79	0.49	0.32	0.15	0.03	0.81	0.14	0.04	0.01
80+	0.55	0.29	0.13	0.03	0.84	0.12	0.03	0.01

If one is interested in comparing the proportions of the population with specific educational attainment across age and sex, then the tabular presentation as given in Table 3 is more appropriate. The table shows that in India for all age groups above 50, more than half of all women are without any formal education. For men this is only true for very old ages (above 70). The table also shows that for primary education, the proportions have become rather similar for younger cohorts. For tertiary education the proportions are highest in the age group 25-29 both for men (with 12 percent having completed tertiary education) and for women (with seven percent having completed tertiary). In the younger age groups the proportions are lower because those cohorts have not yet completed their education. In the older age groups they are lower because of the secular trend of improving education over time. This improvement has been quite pervasive in India, with only five percent of the men and one percent of women having tertiary education in the age group 60-64.

Tables 4 and 5 give comparable information for Egypt and South Africa. On average, Egyptian men and women have higher levels of formal education than their Indian counterparts. While the proportions without any formal education are very high among the older adult population—with more than half of all women above age 45 having no formal education—the proportions with secondary and tertiary education are significantly higher for both men and women. In South Africa, the pattern is quite different (see Table 5). Due to a longer history of primary education for broad segments of the population, the proportion without any formal education never reaches 50 percent even for older women. Actually, the sex differentials are rather small in South Africa. Over the last years there

have been very impressive improvements in education in South Africa which is reflected in the fact that men and women without any education almost disappear in the youngest age groups, and in the age groups below 30 well above 60 percent have completed secondary or higher education. In these age groups women are even somewhat better educated than men. Finally, the data for South Africa also reflect a rather specific African phenomenon where the transition to completed tertiary education tends to happen at rather high ages. For men the proportion with tertiary education only peaks in the age group 30-34. These region-specific differentials in the age at transition to tertiary education will be further discussed in Section 4.5.

Table 4:
Egypt around 2000 (DHS data for 2000). Proportions of the population with four educational attainment categories for men and women by age

	Males				Females			
Age	No education	Primary	Secondary	Tertiary	No education	Primary	Secondary	Tertiary
15-19	0.06	0.12	0.73	0.08	0.19	0.10	0.62	0.09
20-24	0.08	0.13	0.55	0.25	0.22	0.12	0.47	0.19
25-29	0.12	0.15	0.55	0.18	0.33	0.13	0.41	0.13
30-34	0.17	0.18	0.49	0.17	0.39	0.16	0.35	0.11
35-39	0.22	0.22	0.38	0.17	0.48	0.19	0.23	0.10
40-44	0.25	0.25	0.34	0.17	0.48	0.27	0.17	0.08
45-49	0.31	0.26	0.27	0.16	0.56	0.25	0.13	0.07
50-54	0.38	0.25	0.22	0.15	0.64	0.19	0.12	0.05
55-59	0.44	0.23	0.19	0.14	0.70	0.17	0.08	0.06
60-64	0.55	0.19	0.14	0.12	0.74	0.16	0.08	0.03
65+	0.67	0.20	0.07	0.06	0.85	0.12	0.02	0.01

Table 5:
South Africa around 2000 (data from the 2001 census). Proportions of the population with four educational attainment categories for men and women by age

	Males				Females			
Age	No education	Primary	Secondary	Tertiary	No education	Primary	Secondary	Tertiary
15-19	0.03	0.65	0.31	0.01	0.03	0.57	0.39	0.01
20-24	0.06	0.31	0.56	0.06	0.07	0.27	0.59	0.07
25-29	0.08	0.31	0.52	0.09	0.09	0.28	0.52	0.11
30-34	0.10	0.35	0.44	0.11	0.12	0.34	0.42	0.11
35-39	0.13	0.41	0.36	0.10	0.16	0.40	0.34	0.10
40-44	0.16	0.43	0.31	0.10	0.20	0.43	0.28	0.09
45-49	0.20	0.46	0.25	0.09	0.24	0.46	0.22	0.08
50-54	0.24	0.44	0.23	0.09	0.28	0.44	0.21	0.07
55-59	0.26	0.42	0.23	0.09	0.30	0.43	0.20	0.06
60-64	0.33	0.39	0.21	0.08	0.40	0.38	0.17	0.05
65+	0.41	0.33	0.19	0.07	0.49	0.32	0.16	0.04

4.3 Assumptions about mortality differentials and migration

Demographers are aware that mortality rates differ substantially among different socio-economic groups in the population (Kitagawa and Hauser 1973; Preston et al. 1981; Pamuk 1985; Alachkar and Serow 1988; Duleep 1989; Feldman et al. 1989; Elo and Preston 1996; Rogot et al. 1992; Pappas et al. 1993; Huisman et al. 2004). Since a more detailed, direct measurement of these differentials can best be conducted in countries where there is a population register, much of the empirical analysis in this field tends to come from the Nordic countries. Andersen (1991) presented a comprehensive analysis of mortality by occupational status for five countries, Denmark, Finland, Iceland, Norway and Sweden, in which he found, for example, that the standardized mortality rates for workers in hotels, restaurants and on ships is more than two times higher than that of teachers. While occupations can change during a lifetime, the highest educational attainment tends to be a very stable characteristic and is hence very appropriate for the study of socio-economic mortality differentials. In countries that do not have full population registers that automatically give the socio-economic characteristics of every deceased person, so-called matching studies linking the death certificates to the person's characteristics in the previous census can help to obtain the desired information. Doblhammer (1997) found that in Austria, men with only basic education had more than twice the mortality risk of those with tertiary education. For women, the differential is weaker (up to 60 percent) and more pronounced at higher ages (see also Lutz et al. 1999).

Because the direct measurement of mortality by level of education requires a reliable and comprehensive death registration system, together with information on the education of the deceased and the corresponding risk populations, such empirical data are limited to a few industrialized countries and are virtually absent from the developing world. For developing countries the general mortality levels are often estimated from the levels of child mortality that are measured in surveys such as the DHS. Some of these surveys also have information on the number of surviving relatives from which one can infer information about adult mortality. While such procedures can provide useful estimates for the levels of overall mortality using model life tables for total life expectancy, they do not allow us to estimate education-specific mortality levels because typically only the education of the respondent in the survey is known and not that of the deceased relative. Hence neither direct registration of deaths nor inference from surveys can help us gain such information for a large number of developing countries. This leaves us with only the third piece of information that is usually available for most countries, namely, a sequence of censuses.

If one has a series of at least two censuses, e.g., for Kenya in 1989 and 1999, which are both considered to be fairly reliable and give the total population by age, sex and level of educational attainment (in comparable categories), one can quite easily calculate census survival ratios, i.e., compare the number of women

without any education aged 45-49 in 1989 to the same category of women aged 55-59 in 1999. If women have not gained further education, i.e., moved educational categories between the ages 45 and 59, then the ratio of the two sizes of this same cohort gives a combined estimate of education-specific survival and net migration for the age groups concerned. In order to obtain a rough estimate of this kind of education-specific census survival, we carried out such an exercise for Brazil, China, France, Kenya, Malawi, Mexico, Uganda and Vietnam. We examined the survival of cohorts of people aged 40-49 in each educational category over 20 or 30 years through three to four decennial censuses for several countries of the world, as permitted by data availability. The choice of this age group was motivated by two competing objectives: The older the studied cohorts are, the higher is the chance that they will not experience further changes in educational attainment status; but if the cohorts chosen are too old, there is a higher chance of age misreporting and the danger that the cell sizes will be too small. Our sources were tables from census reports and data from Integrated Public Use Microdata Series (IPUMS) (http://www.ipums.umn.edu/). Under the IPUMS program, massive amounts of micro-data from national census samples are now becoming available. Within a few years nearly 200 of these samples will be available covering over 50 countries. This growth in the availability of census samples will allow us in the future to investigate the dynamics of changes in the educational composition of many populations in detail that previously would have been impossible to attain. However, for the time being we had to limit our analysis to the eight countries mentioned. This extensive exercise was carried out at IIASA in 2005 and the findings were reported in separate papers (Sanderson 2005; Fotso 2006; Woubalem 2006; Figoli 2006) and shall not be described here in any detail.

For several reasons we decided to capture the educational mortality differentials in terms of life expectancy at age 15 (*e15*). The life expectancy at birth includes the infant and child mortality experience, which also depends on the educational level of the parents, but this is not what we want to measure. Further, lifetime educational attainment of an individual might not affect survival in lower ages. We assumed that the effect of the education of an individual on mortality starts at around age 15. Around this age people start to join the labour force and the type of job they get is usually related to their current educational attainment at that age and to some extent their expected future educational attainment.

For the countries studied, we found an average increase in *e15* of one year from the no education category to the primary education category. In contrast, we found an average increase in *e15* of two years from the primary education category to the secondary education category and also from the secondary education category to the tertiary education category. It is interesting to note that practically all of the countries studied showed this pattern of a smaller differential between the lowest two categories. Also, this pattern of two years difference in

life expectancy between the highest categories fits well with the general pattern of educational mortality differentials directly measured in some of the industrialized countries, as discussed above. For instance, a recent, very detailed study from the Swedish population register shows that in the year 2000, life expectancy at birth for men with nine or less years of education was 75.8 years, for 10-11 years of education 77.0 years and for 12 or more years of education 79.3 years, while men with higher academic training are expected to live more than 80 years (Batljan, work in progress). This implies that even in countries with very low mortality, the differential among the lower education groups is smaller than among the higher.

Assuming for the time being that this pattern of a one-year differential in *e15* between the two lowest categories and a two-year differential each between the other categories holds for all countries and for the entire period 1970-2000, how should this be operationalized in our back projections? If we know *e15* for any specific educational category in a country, we could then use these educational differentials to obtain *e15* for each educational category. In our study, we use the population data produced by the UN Population Division (UN 2005) for all purposes. We used the same source to obtain *e15* for the total population of all countries for each five-year period from 1970-1975 up to 1995-2000. This is not sufficient, however, because in order to apply the differentials, we need *e15* for each of the educational categories, which is not given by the UN or any other source. To solve this problem, we decided to anchor the population life expectancy to one educational category. To do this, we need to choose a category that has a high proportion of the total population. Choosing tertiary or no education would not make sense, since they are two extreme categories with very few people at both ends of the development spectrum. The choice was thus between primary and secondary. We chose secondary because, on the global level, this seems to be the most rapidly expanding category. Alternatively, we could have had different anchor categories in different countries and at different times, but this would have added an unnecessary further level of complexity.

Using the UN (2005) dataset and the general UN model life table, we find the *e15* for every country and for every period. This gives us the *e15* of the population that is a weighted average of *e15* for each educational category. We then assume that the *e15* given for the total population will be approximately equal to the *e15* for the secondary category. Based on this assumption we are now in the position to apply the educational differentials in *e15* and produce estimates for the mortality levels in all educational categories. If left uncorrected, this procedure will lead to an upward bias in the overall level of life expectancy in countries where more people are in categories above secondary than below secondary, and to a downward bias in countries where the opposite is the case, i.e., in poorly educated, developing countries. However, in our procedure outlined in Box 1, a somewhat distorted level of overall life expectancy at this step in the back projections is of no consequence because the resulting total age structure will be readjusted proportionately to exactly match the age structure given by the

UN (2005), thus automatically applying the right overall level of mortality. Hence, only the relative mortality differentials matter for the reconstruction of the proportions in different educational categories. These relative differentials remain unaltered throughout this anchoring procedure.

Finally, we will briefly discuss migration. As discussed above, changes along cohort lines can only be caused by mortality, by migration if we consider the total population and by changes from one educational category to another if we consider education-specific cohorts. These educational transitions will be discussed in Section 4.5. Mortality has already been discussed above. There is no easy way of dealing with the only remaining factor, migration, because unlike mortality there are no systematic differentials. In some cases, depending on the specific nature of migration, immigrants have a lower educational profile than the receiving population while in others they have a higher one. Hence, for migration we cannot estimate a typical differential but rather have to worry about each individual case.

But although unsatisfactory, this limitation is not very serious for the following reasons: it is important to first understand that we do not have to worry about the total volume of migration, just like we do not have to worry about the overall level of mortality, because the adjustment to the UN population structure will take care of this. The only thing we have to worry about is the case in which migration significantly alters the educational composition of the population. This is clearly not the case when the educational composition of net migration is equal or similar to the educational composition of the population under consideration. There is only reason to worry if (a) there is a significant level of net migration (either migration gain or loss), (b) the educational profile of this gain or loss is significantly different from that of the resident population, and (c) the age pattern of migration is rather old so that it affects several age groups in the back projections.

Let us consider the three criteria separately. For (a), the UN data give estimates of the total volume of net migration although they are mostly derived as residuals once birth and death rates are given. As to (b), there is no empirical data on migration by level of education for most countries in the world. Hence, there is little that can be done to assess this criterion. Concerning (c), it can be said that migration usually happens at rather young ages with typical migration profiles showing a peak in the age group 20-24 and a second smaller peak in the age group 0-4 for migrants arriving with their children. While the migration of children can be safely disregarded in this model, migration in the age group 20-24 can only affect the reconstructions of the age groups 15-19 and 20-24, which are already somewhat problematic because of the assumptions that have to be made on the age of transition to higher categories, as will be discussed in Section 4.6. If no migrants arrive beyond the age of 25, this will not affect the estimates for all age groups above 25, because when going backward in time, we move from the older age groups (that already reflect past migration) to the younger ones. Hence,

one can assume that for the majority of countries, migration will not present a major distorting force. But there are a few countries—Israel is probably the most extreme case—where all three criteria are met and our reconstruction is likely to be biased. In such cases the only solution is to correct the reconstructed data through empirical data, if they are available (as will be discussed in the validation section below) or otherwise not include such countries in the dataset.

4.4 Dealing with the open-ended age group

One problem that is common to all back-projection efforts is the fact that in all empirical datasets, the highest age group is usually an open one, such as 65+ as is the standard in our baseline data. Some countries have more information about the older age groups, such as India which has information up to the age group 80+ (see Table 3). We took advantage of this information whenever we could. Therefore, the procedure described below did not have to be applied or was only applied for the years 1970-1985. This problem with the open ended age group was also the main reason why we stopped the reconstruction in 1970. For instance in the case of an open-ended age group of 65+ in 2000, it would translate through the reconstruction into an open-ended age group of 35+ in 1970. However, if we go further back, this would lead to an additional increase in the proportion of the data estimated by extrapolation rather than on real data.

At every back-projection step, the task is to estimate from the given open interval 65+ the proportions in different educational attainment groups for the age group 65-69 which, after this step, will become the age group 60-64. The following procedure will describe how we estimate these proportions for the 65-69 age group based on the extrapolation of the trend as derived from the proportions in the younger age groups. This procedure is done in several iterations to make sure the estimates are consistent with the known education proportions for the highest open age group (65+). While doing so, we also consider that proportions always lie between 0 and 1, and that the sum of the proportions in each age group must equal unity.

Let *age* = 1, 2, … , 10 represent the age groups 15-19, 20-24, … , 60-64, *educ* = 0, 1, 2, 3 represent educational attainment levels, namely, no education, primary, secondary and tertiary and let *y(age,educ,t,sex)* represent proportions of people in age group *age* with education level *educ* in a given year, *t*, separately for males and females.

By definition, $\sum_{educ} y(age, educ, t, sex) = 1$.

Let $educ'$ be a specific level of education. The proportion of people of a given age and gender in the year 2000 who have at least that level of education can be written as:

$$Y(age, educ', 2000, sex) = \sum_{educ=educ'}^{3} y(age, educ, 2000, sex) \quad (2)$$

The educational attainment progression ratio, $EAPR(age, educ', 2000, sex)$, is the proportion of people of a given gender in 2000 who have education levels higher than $educ'$ among those with education level $educ'$ and higher. We write:

$$EAPR(age, educ', 2000, sex) = \frac{Y(age, educ'+1, 2000, sex)}{Y(age, educ', 2000, sex)}. \quad (3)$$

Note that *EAPR(age,3,t,sex)*=0 for all years.

The value of $EAPR(age, educ', 2000, sex)$ cannot be less than zero or greater than 1, as both numerator and denominator are non-negative and $Y(age, educ', t, sex) \geq Y(age, educ'+1, t, sex)$. To ensure that the *EAPRs* were always in this range, we worked with the *logit* of *EAPRs* in analysis. If we assume that the logit has two advantages, first, it will make sure that the EAPR always remains between 0 and 1. Secondly, when plotted on the graphs, logits of EAPRs show a higher degree of linearity in most cases. The property of the logistic curve being asymptotic to the boundaries (ex. EAPR to primary = 1) represents well the behaviour of EAPR's, with a slow increase at the beginning, followed by a faster increase and then slowing down again towards the end.

For each education and gender group, we ran the following linear regression:

$$\text{logit}[EAPR(age, educ', 2000, sex)] = a(educ', sex) + b(educ', sex) \cdot age + \varepsilon, \quad (4)$$

for *age=6, 7, ... , 10* (which represent the age groups 40-44 up to 60-64).

The estimated coefficients were used to extrapolate for five-year age groups between 65 and 100 and unfolded to obtain $y(age, educ', 2000, sex)$ for age groups 65-69 through 95-99 (*age=11, 12, ... , 17)*. In other words, this procedure extrapolates the proportions in the age groups based on the trend of changes—mostly improvements—that is observed for up to five older cohorts.

Next, we have to ensure that the extrapolated education proportions add up to the original education proportion of 65+. We do this by adjusting the constant term in Equation (4) so that the difference between the education proportion of 65+ obtained from extrapolation and the original is insignificant ($<10^{-6}$).

This procedure was applied at every step in the back-projection process unless empirical information for the higher age groups was available. What usually

happens when going back in time is that the trend is derived increasingly from proportions that were estimated themselves. In order to avoid this, we progressively lowered the ages that were included in the regression.

Finally, it is worth noting that an analogous procedure was used for interpolating the base year data when the information was not provided in five-year age groups but rather in broader age groups.

4.5 Age at progressing to higher attainment category

Changes in educational attainment by age and sex follow a hierarchical multi-state model, which implies that transitions from one educational category to another can only go in one direction and have to follow a predefined sequence. This means that people over time can only move to the next higher educational attainment category step by step and cannot move backward. Somebody who has once reached completed tertiary education will maintain this status throughout his/her life no matter what happens to the person's actual skills or abilities. This follows from the definition of a formal level of educational attainment chosen here, which is the only approach possible given the nature of the empirical data. Should more systematic information on actual skills by age become available for several points in time, one could also think of applying models that explicitly capture the possible deterioration of skills.

In the case of forward projections, it is both the timing and the quantum of transitions that matters. In the case of back projections, the quantum, i.e., the ultimate proportion ending up in a certain educational category, e.g., by age 40, is given. If every person has completed his/her education when reaching the age group 40-44, we know from the given data for this age group in 2000 what proportion of this cohort will end up in the different educational categories. When going back in time and to younger age groups of this same cohort, the only question that we need to worry about is the timing of transition. In other words, we need to estimate how many of those who have tertiary education in the age group 40-44 in 2000 already had tertiary education in the age group 20-24 in 1980.

Since in this reconstruction effort we only go down to the 15-19 age group, as the lowest age group for which we reconstruct the data, transitions that typically happen before this age need not be of concern here. This is clearly the case for the transition from the category no formal education (E1) to that with some primary education (E2). But the issue already becomes more problematic for transitions from primary (E2) to the completed lower secondary education (E3) and completed tertiary categories, where a certain proportion is expected to still happen between ages 15 and 19. The transitions to completed tertiary (E4) clearly can happen in a broad range of age groups. While the timing of transitions to E3 will only require some assumptions about the age group 15-19, the transitions to tertiary clearly require more consideration. The main problem is that the ages at

transitions to E4 vary greatly among countries. For example, before 1997 the Bachelor's degree in Nepal only took two years and many people finished it at the age of 20. In contrast, in some African countries, it is not uncommon to receive the first university degree after the age of 40. For this reason we need some country-specific assumptions for the transitions to E4.

When discussing the assumptions concerning the ages of educational transition, it is useful to refer to Table 2 which gives a detailed account of the nine educational categories that have been collapsed into the four categories used in this exercise. Although the duration of these nine categories show considerable variation across countries, it is still useful to consider this finer classification when defining the assumptions made.

The transition to completed lower secondary education (E3) typically happens around the age of 14 in most countries. However, due to reasons of late entry, repetition and in some countries to a longer official duration of lower secondary education the proportion of a cohort making transitions to E3 after the age of 15 varies greatly among countries. For this reason we had to make country-specific assumptions concerning the proportions of transitions to E3 after age 15. We did this for a given year by taking the difference in proportion of those in E2 in age groups 15-19 and 20-24 (as an estimate of those making later transitions to E3) and divided this difference by the actual proportion in E3 in the age-group 20-24. This then gives us the proportion of late movers among all movers to E3. To illustrate this, in the Comoros in 2000 for males the proportions E2(15-19) and E2(20-24) are 0.56 and 0.36 respectively, the difference hence being 0.2; the proportion E3(20-24) is 0.41. The proportion of late movers is therefore 0.20/0.41 = 48%. Though only a rough approximation, this is the best country-specific estimate that can be derived from the given data. In the back projections it will only affect the proportions in E2 and E3 in the age group 15-19.

Typically, most of the transitions to the tertiary educational attainment category that occur by completing the first level of tertiary (generally called a Bachelor's degree) happen around the age of 22. However, due to repetition and late or delayed entry into the education system, significant proportions of people complete the first level of tertiary at later ages. As mentioned above, these transitions show great variations across different cultures and countries and therefore, we use the following country-specific strategy: First, we look at the tertiary category in 2000 across the age groups to find the age group with the highest proportion. If the peak is in the age group 20-24, we assume that all the people who had tertiary education in the age group 20-24 had secondary education in the age group 15-19. If the peak is in the age group 25-29, we assume that two-thirds of those who have tertiary education in this age group were in the secondary category at age 20-24, i.e., had not yet completed their education at this age. By the age group 15-19, all are assumed to be in the secondary category. If the peak turns out to be in the age group 30-34, we assume that one-third of them were still in the secondary category at age 25-29, two-

thirds were in the secondary category at age 20-24 and all of them were in the secondary category at age 15-19.

Two different forces can cause peaks in the proportions tertiary: The completion of the transitions of a cohort at a certain age and a time trend in which different cohorts have different levels of tertiary completion. If the proportions reaching tertiary education have an increasing tendency over time—as is the case in most countries—this will lead to a peak at younger ages. If the peak is beyond the age of 35—an age when most people typically have reached their maximum education—this can also be taken as an indication that educational conditions have been deteriorating over time, as is the case in some African countries. Hence, if the peak is beyond age 35, we assume that there was a decline in the education transitions to level E4 over the past decades.

4.6 Conversion of attainment distribution into mean years of schooling

As mentioned in the introductory discussion, many analyses of the implications of changing levels of educational attainment prefer to use just one average indicator of human capital rather than the full distribution across educational attainment categories. Presumably they prefer this less informative indicator because it is simpler to use only one number as a human capital indicator in the various regression models. However, this one number not only hides the interesting educational attainment distribution, it is also a more problematic indicator than the distribution, because it is never measured directly but has to be derived from the attainment distribution by applying even more problematic assumptions than is the case for reconstructing the distributions themselves. Because this summary indicator of mean years of schooling is necessary for comparisons to the many studies that choose to use only this indicator, we decided, as a further derivative of our reconstruction, to produce this simple indicator based on making additional, not unproblematic assumptions. However, we make it simultaneously clear that whenever possible, the full distribution should be used for analysis.

Unlike most other reconstruction efforts, we divide the age structure of human capital into five-year age groups. This suggests that we are not only in the position to produce the mean years of schooling for the entire adult population—as most authors do—but also calculate age-specific mean years of schooling. In the output tables from our calculations (see Table 6), these age-specific summary measures are given as the right margin of the matrices, with the mean years of schooling of the entire adult population being the figure trying to summarize the whole matrix in the bottom right corner.

Barro and Lee (1996) calculated mean years of schooling by using their data on the distribution of educational attainment among the adult population and the information for each country on the time it takes to reach each educational level. Cohen and Soto (2007) do not mention how they calculated the mean years of

schooling. De la Fuente and Doménech (2006) estimated the average years of total schooling by using assumed cumulative durations of the different levels of schooling. They also note that their results are not directly comparable with the Barro and Lee average schooling series.

We calculated mean years of schooling using the time it takes to reach each educational level according to data from UIS for each country, and combining this with our data on the population distributed by age, sex and educational attainment levels. However, these data cannot be directly converted into mean years of schooling for our four categories for several reasons: First, we have broader aggregate categories. For example, the secondary educational attainment category consists of people who have either completed lower secondary (ISCED2) or upper secondary (ISCED3), or have incomplete tertiary. If we used the total duration for completed secondary, we would overestimate the mean years of schooling in our category E3. Similarly, primary educational attainment consists of people who have incomplete primary, complete primary (ISCED1) and incomplete lower secondary. Lastly, tertiary educational attainment consists of people with completed first level of tertiary (ISCED5), incomplete second level of tertiary (ISCED6) and completed second level of tertiary (ISCED6). Hence, some corrections have to be made to arrive at an estimate for mean years of schooling that is representative of the actual mean years of schooling for people in our categories.

We made the following assumptions: For the primary category (E2) we assumed three-quarters of the duration it takes to complete primary as the mean years of schooling for this category to account for those who have incomplete primary. For category E3 we took half of the duration of schooling in upper secondary and added this to the cumulative duration of schooling to completed lower secondary to arrive at an estimate of mean years of schooling in our data. Finally, for category E4 we took half of the duration of schooling in lower tertiary and added this to the cumulative duration of schooling to completed lower tertiary to estimate the mean years of schooling for those who attained tertiary education in our data. These assumptions seem rather crude and we acknowledge that we could be more careful in calculating mean years of schooling, but this would entail a great deal of effort. A more careful analysis would have to address many difficult issues affecting mean years of schooling, such as how to deal with repeaters, dropouts, hours of schooling per day, days of schooling per year, etc. This type of more careful analysis will have to be looked at in the future. But by being very explicit about the rather simple rules which we applied to all countries, we are already doing better than several of the other datasets that leave the users in the dark about how exactly they derived their mean years of schooling. We consider the reconstructed age-specific proportions in the different educational categories as our main product, and whenever possible these data should be used instead of the further derived mean years of schooling.

5 Selected results

This reconstruction exercise resulted in an unprecedented amount of detailed and consistent information for levels of education by age and sex for 120 countries over three decades. Our database contains this information for five-year age groups and in five-year time steps from 1970 to 2000 for 120 countries. This database is openly available (as of the end of 2007 on the web site of IIASA's World Population Program, http://www.iiasa.ac.at/Research/POP/edu07/index.html as well as the VID web site, http://www.oeaw.ac.at/vid). The standard output file for each country consists of four parts (typically displayed on four A4 pages). Owing to space limitations we can only give this full set for one country (see Appendix A). We chose India, which recently joined IIASA as a member country, as it shows a very interesting pattern of improving human capital, and will soon be the world's most populous country, surpassing China in population numbers but not in terms of human capital.

The first of these four standard output pages for every country (as well as for major world regions and all 120 countries taken together) gives the set of seven multi-state age pyramids in the same form as presented for Egypt in Figure 1. It shows the empirical age pyramid by level of education for 2000 containing all the empirical information needed for the reconstruction exercise, followed by age pyramids for 1995 to 1970 in five-year intervals. These multi-state pyramids present very useful visual summaries that allow the reader to catch the main patterns at a glance, something that is not so easy from the massive numerical information presented in the other three pages. The second page shows the full series of age and education matrices in five-year steps for the absolute numbers of men and women in each category, while the third page gives the same information for proportions. The fourth page shows the absolute numbers and proportions for men and women combined. Together, these three pages of tables provide about all the information that a potential user might possibly need.

Table 6 gives an example of one of these standard matrices for Egyptian women in 1980. Each of these matrices has the age dimension along the vertical axis in five-year age groups from the population aged 15-19 to the highest category 65+. At the bottom, it lists the sum of the population 15+. In order to make it directly comparable to the Barro and Lee data, we also included the aggregate age group 25+. Along the horizontal axes, it lists the four educational attainment categories considered. At the right margin, as a summary measure across educational categories, the matrix lists the age-specific MYS calculated according to the procedure described in Section 4.6. In the lower right corner, we find the summary measure along both dimensions that is the mean years of schooling for the entire population above age 15 (and above age 25). In this specific example of Egyptian women, the pattern of age-specific mean years of schooling is quite revealing. It shows significant educational improvements that had already happened in the decades preceding 1980. While women aged 15-19 in

1980 had on average 4.2 years of schooling, it monotonically decreases with age and reaches a low 0.6 mean years of schooling for those aged 60-64. The matrix also shows that for all women above the age of 25, more than half have never been to school. This proportion reaches 90 percent for women aged 60-64 in 1980.

Table 6:
Standard output matrix for the absolute number of Egyptian women in the year 1980

Females		No Edu.	Primary	Secondary	Tertiary	MYS
1980	15-19	1056.7	451.1	695.0	0.0	**4.2**
	20-24	972.4	561.6	444.5	19.1	**3.8**
	25-29	946.7	440.4	254.6	39.6	**3.2**
	30-34	840.1	260.2	164.3	32.9	**2.7**
	35-39	714.8	189.2	86.5	29.5	**2.2**
	40-44	728.7	159.0	77.9	13.6	**1.8**
	45-49	722.5	121.7	54.8	9.3	**1.4**
	50-54	653.6	84.7	34.9	5.7	**1.1**
	55-59	557.8	55.4	20.7	3.3	**0.8**
	60-64	449.0	34.0	11.5	1.7	**0.6**
	65+	898.1	40.9	11.6	1.6	**0.3**
	15+	*8540.4*	*2398.3*	*1856.5*	*156.4*	***2.5***
	25+	*6511.3*	*1385.6*	*717.0*	*137.3*	***1.8***

Since this is not the place to discuss many country-specific results in any detail, we will use the rest of this section to look at the global summary of all 120 countries together and then compare the two demographic billionaires, China and India. Figures 2, 3 and 4 each consist of two parts: (a) gives the trends in the absolute size of the number of men and women combined in the specified age range; (b) gives the changing proportions. As we will see, these two different ways of looking at the data may suggest quite different interpretations.

Figure 2 gives the trends for all 120 countries taken together. These make up 93 percent of the world total. Hence, we may safely call it the global trend. Figure 2a illustrates the tremendous expansion of the global working-age population, which was around 2 billion in 1970 and almost doubled to more than 3.5 billion in 2000. It also shows that the size of the lower two educational categories hardly changed over time. There are almost as many uneducated people of working age in the world today as there were in 1970. Some people may interpret this as a spectacular failure of educational policies that, despite all efforts, did not manage to lower the number of uneducated people on this planet. However, the picture looks very different when we look at the change in proportions (Figure 2b), which is dominated by a rather spectacular expansion of the people of working age with completed secondary education. The proportion of uneducated people in this view has strongly declined from more than 35 percent to less than 20 percent in only 30

Figure 2a:
Population distribution by education for population aged 15-64 in 1970-2000 in the world (120 countries/economies)

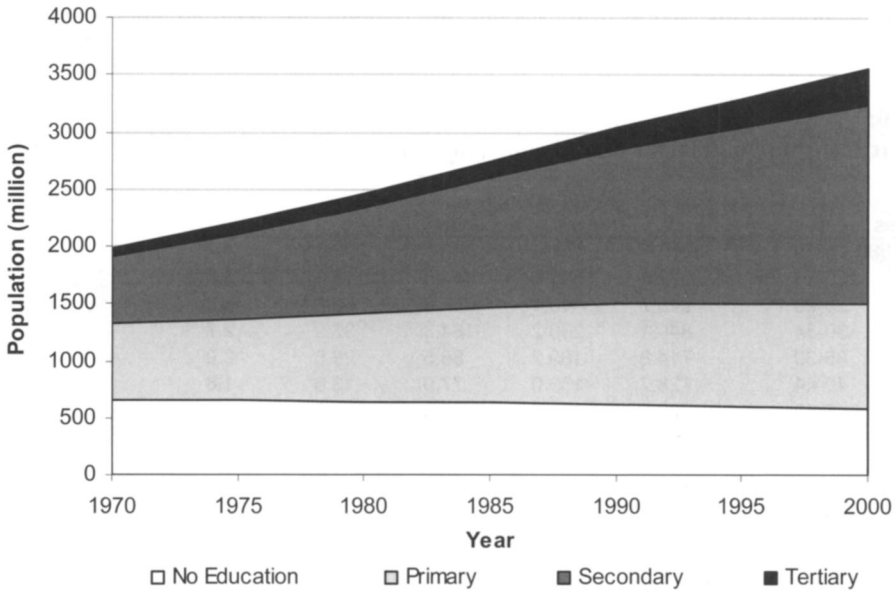
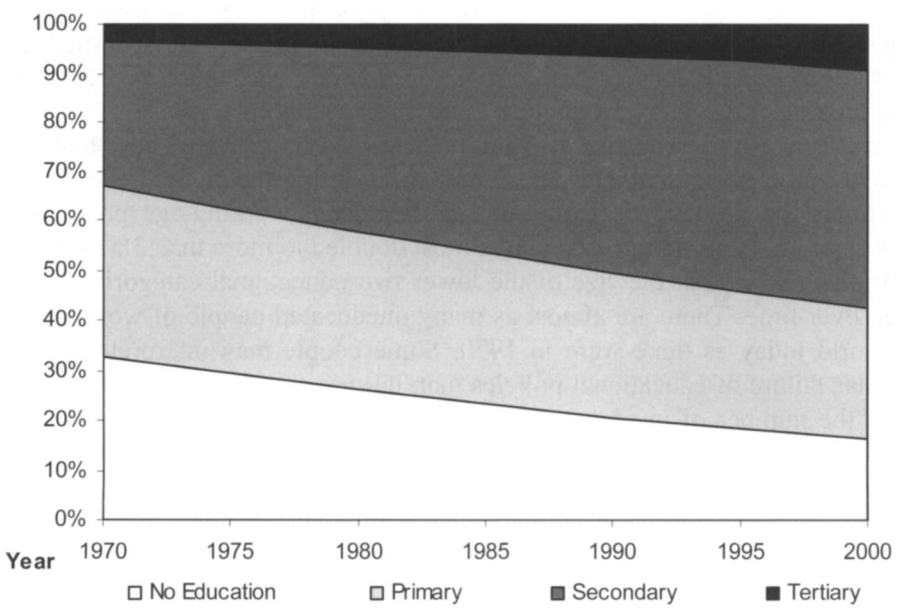


Figure 2b:
Proportion distribution by education for population aged 15-64 in 1970-2000 in the world (120 countries/economies)



years, which sounds like a success story. This dependence of the story on the perspective taken is very similar to the discussions around poverty eradication, where some groups point at the fact that the number of people in poverty has hardly declined over time, while others stress the fact that the proportion of poor in the world has dramatically declined. Actually, the trends in the prevalence of poverty may be very closely correlated to the trend in the prevalence of uneducated people, something that can now be studied more comprehensively using our new data, which includes the distributional dimension.

Figures 3 and 4 show similar graphs for the demographic giants, India and China. In order to focus our attention on the more recent improvements in educational attainment, we chose not to look at the entire working-age population, but only at the population aged 20-39, the age groups that are presumably key to social and economic innovation. In India, the trend in absolute numbers looks somewhat worse than on the global level discussed above. While the young adult population more than doubled over the 30 years from 1970 to 2000, the number of people without any formal education increased from 93 million in 1970 to 117 million in 2000. The number of people with primary education only also expanded over time. This seems to imply that the educational system, despite major efforts, could not keep pace with population growth, and the lack of education is a bigger problem today than it was in the past. Again a look at changing proportions in Figure 2b brings out a much more favourable picture of improving proportions over time.

The Indian trend (Figure 3) is in great contrast to that in China (Figure 4). While in China the young adult population also doubled between 1970 and 2000, the educational system was evidently much more effective and managed to reduce the number of uneducated persons very significantly, even in absolute numbers. Also, the number of people with only primary education declined over that period. This was associated with a most spectacular increase in the number of young adults with secondary education. It increased six-fold from 50 million in 1970 to 300 million in 2000. This spectacular increase in educational attainment in the world's most populous country is a fact that has hardly been acknowledged in the discussions about the recent rise of China on the world stage. Not surprisingly, these improvements in the Chinese human capital stock look even more impressive on the relative scale (Figure 4b). The remarkable difference between India and China in terms of human capital should be kept in mind when considering the global impact and the likely future economic power of these two giants.

Figure 3a:
Population distribution by education for population aged 20-39 in 1970-2000 in India

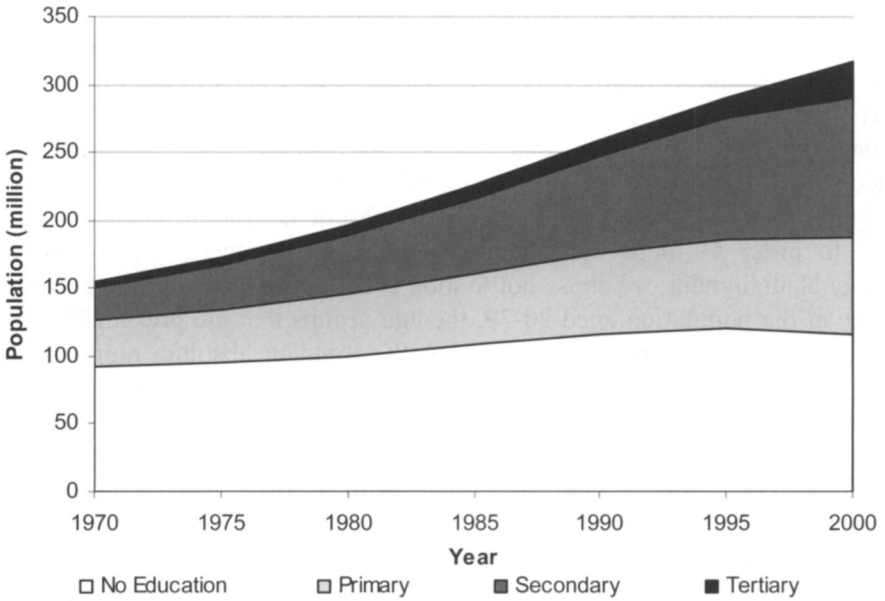


Figure 3b:
Proportion distribution by education for population aged 20-39 in 1970-2000 in India

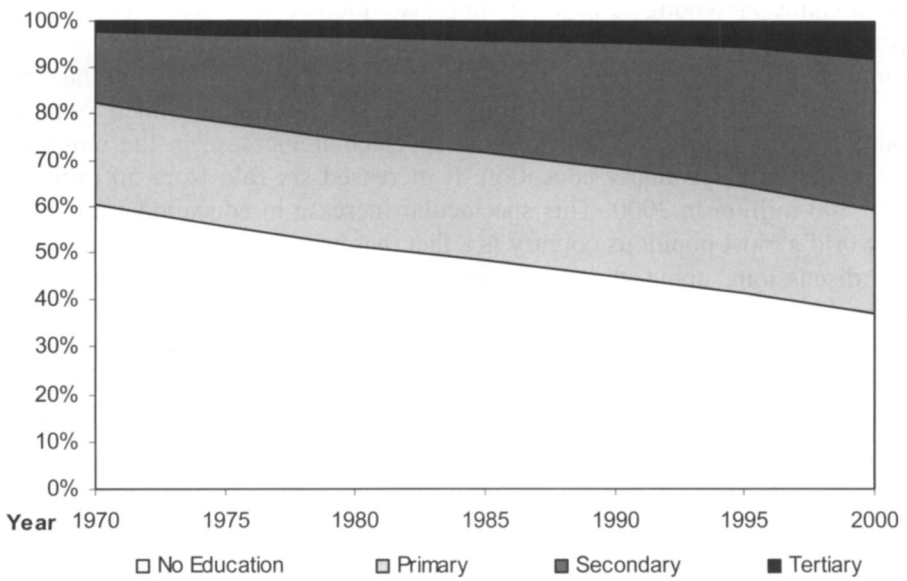


Figure 4a:
Population distribution by education for population aged 20-39 in 1970-2000 in China, Macao and Hong Kong

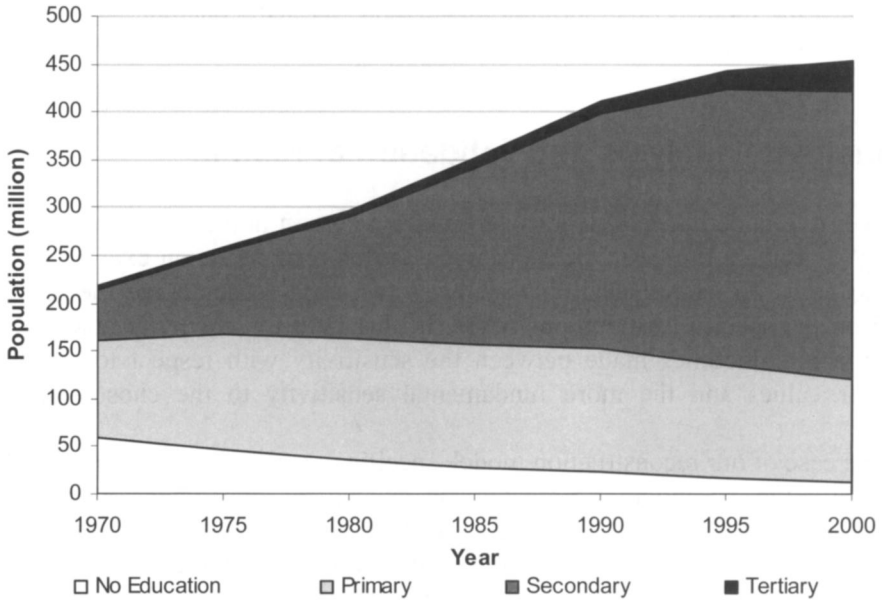
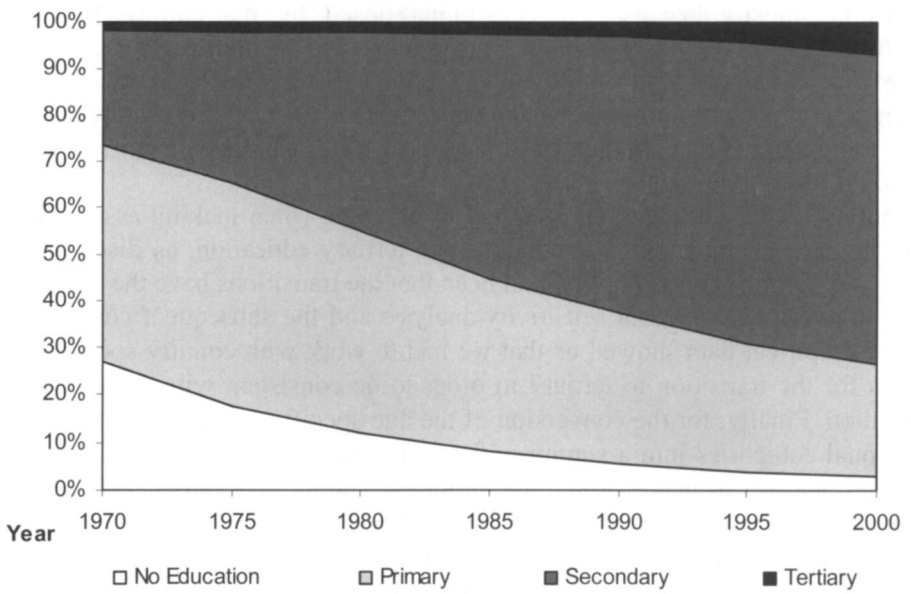


Figure 4b:
Proportion distribution by education for population aged 20-39 in 1970-2000 in China, Macao and Hong Kong



As indicated earlier, this paper will not enter a systematic discussion of the interpretation and implications of the newly reconstructed dataset. There will be several papers in the future that will substantively study these reconstructed trends. Here the focus is on describing the method. We will conclude with a brief discussion on sensitivity analysis.

6 Sensitivity analysis and validation of results

Whenever one undertakes analyses that require a certain number of assumptions in order to produce the desired results, such as this reconstruction exercise, it is useful to carry out some sensitivity analyses to see to what degree the results depend on the specific assumptions made. In this field of sensitivity analysis, a distinction is sometimes made between the sensitivity with respect to specific parameter values and the more fundamental sensitivity to the chosen model structure.

In the case of our reconstruction model, specific parameter assumptions had to be made for the extrapolation procedure in closing the open-ended interval as described in Section 4.4. While there was little choice in terms of using an extrapolative procedure to produce some proportions (which are each constrained to lie between zero and unity and which sum up to unity), there was some degree of freedom in terms of the number of younger age groups which serve as the basis for extrapolation. We experimented with different lengths of the reference age groups and found only very minor differences in terms of the estimated proportions, mostly because the constraints posed by the empirically-given proportions for the entire open-ended age group were so dominating. The final choice to include the five empirical age groups before the age group to be estimated was a compromise between the objective to give more emphasis on recent trends and the contradictory objective to have a broader empirical input that would result in more stable estimates.

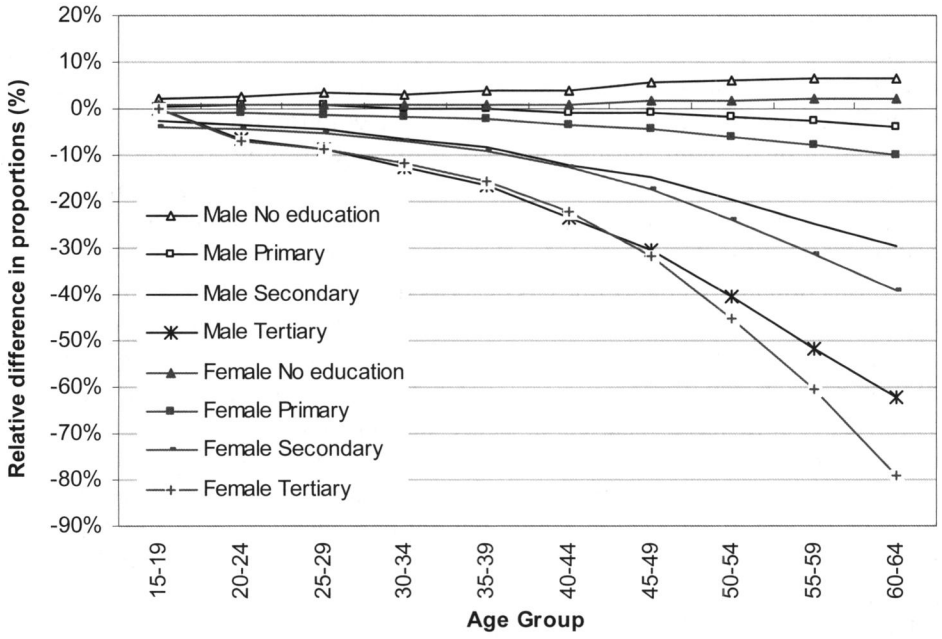
Another set of parameter choices had to be made when making assumptions about the ages at transition to secondary and tertiary education, as discussed in Section 4.5. Our initial assumption had been that the transitions have the same age profile in every country. But sensitivity analysis and the subsequent comparison with the empirical data showed us that we had to work with country-specific age patterns for the transition to tertiary in order to be consistent with the empirical distribution. Finally, for the conversion of the age-specific proportions in the four educational categories into a summary figure for the age-specific mean years of schooling, we had to make certain assumptions. The chosen values were argued extensively in Section 4.6. We also conducted a sensitivity analysis which was rather simple, because the effect on the output (mean years of schooling) is simply a linear function of the specific values chosen for mean years of schooling

in each attainment category. The chosen values are those considered to be most plausible and defendable.

In terms of structural assumptions, the only choice that falls into this category seems to be the consideration of educational mortality and (implicitly) migration differentials. Otherwise, the model of population dynamics along cohort lines is unambiguously established as discussed above and has no real alternative candidates to be tested against. The only alternative might have been to go more in the direction of Barro and Lee's work and use empirical school enrolment rates as an additional input influencing not only the timing of the educational transitions but also their quantum, i.e., the transitions to higher attainment levels. In this case, however, our model would have been over-identified because the final attainment levels are given by the empirical distributions for 2000 and our back projection along cohort lines as described in this paper. Hence, incorporating empirical data on enrolment rates either would present a confirmation of the transition probabilities derived from our model (this would be expected in the case of perfect data and assumptions for both enrolment and our attainment model) or would have resulted in some discrepancies that would have to be resolved. In case of such conflicts, the transitions derived from enrolment data would, in most cases, have weaker credibility for methodological reasons (they tell us how many people are in school but not how many of them finish a degree) and data collection bias (reported by schools who have a vested interest in showing that they meet certain targets). For this reason, it did not seem to be a viable alternative model to us. But this does not imply that we will completely ignore empirical information on enrolment as part of the validation exercise as discussed below.

One key choice that distinguishes our reconstruction efforts from all the others so far is the explicit consideration of educational mortality differentials. For this reason it seems worthwhile to have a closer look at the sensitivity of the results to this model choice. In order to assess this sensitivity, we performed an alternative reconstruction for a selected number of countries for which we used an otherwise identical model but assumed that all educational attainment groups would be exposed to the average mortality prevalent in the total population at the time period considered. Figure 5 shows the difference between the results of the two models. For each educational attainment category and for men and women separately, the figure shows the relative difference in proportions which was calculated by dividing the difference in proportions resulting from the two models by the level of the proportion resulting from our model that uses educational differentials. The figure compares the results by age for the year 1970 using the example of India.

Figure 5:
Relative difference in proportions in educational categories of our model as compared to an alternative model without educational mortality differentials by education, India, 1970



At first sight, the figure reveals two important features: That differences increase with age and that differences are much more pronounced for the higher educational categories. The increase with age results from the fact that mortality rates are much higher at the older ages and therefore, differentials in the mortality rates affect the number of people in each education category more significantly. When discussing the different impact on the four educational attainment categories, this has to be seen in relation to the average level of education in the population which was very low in India over this period. Only the no-education category was slightly below the national average and hence had higher mortality than the average in our model, which results in a larger number of people without education when going backward in time, i.e., adding the people that we assume have died to the size of the cohort. For the three other education groups, the opposite is true. By assuming lower than average mortality for the better educated groups, we will add less people to cohorts when going back in time, which will result in lower estimates of the sizes of the educated groups in 1970. As we see from the graph for the highest educational group, this relative difference reaches 60-80 percent fewer men and women with tertiary education than we would obtain when disregarding the educational mortality differentials. Although this is the cumulative effect of the reconstruction over 30 years, this still impressively

demonstrates that considering the mortality differentials explicitly does indeed make a significant difference.

Finally, one activity that has only started under the project described here is the validation of our reconstructed results against all the empirical data that are given by old censuses (mostly from the UIS database) as well as older surveys and national series of school enrolment rates at different levels. As mentioned in the introduction, we are presenting here Version 1 of our dataset, which gives the data as reconstructed and subjected to a first round of validation. In this first round we compared our reconstructed results to the historical data given in the UIS database and other data we had received directly from the national statistical agencies. In this first round of validation we applied two clear criteria to identify significant discrepancies: If our reconstructed proportions, at any level, age group or point in time, deviated by more than five percentage points or by more than 20 percent on a relative scale from the other data source, it was classified as an outlier that needed further attention. We then made an in-depth analysis for all the outliers to try to determine the source of the discrepancy. In many cases we could resolve the problem either by finding that the definition of educational categories differed in the other source (the most common problem) or that our assumption of no significant education differentials in migration was violated and we could make a plausible correction of this assumption. A handful of cases remained unresolved and since the discrepancies were significant, we decided to remove these countries from our dataset. These countries are not part of the 120 countries presented here, which still represent 93 percent of the world population.

For the future we foresee more detailed validation exercises in direct collaboration with the UIS. We will not be satisfied with the stated tolerance limits, but will try to resolve all discrepancies so that in the end, a corrected and completed (based on comparison to our reconstruction) UIS historical dataset and our further validated reconstruction dataset become identical. In this process we will also rely on all available time series on school enrolment rates. This will be a major effort which is likely to take about two years and will result in a second version of the dataset.

7 Conclusions and outlook

This paper gave an overview of the demographic back projection method that was used to estimate a new comprehensive and detailed dataset on human capital by age and sex. Together with the first round of the validation exercise as described in the previous section, this constitutes Version 1 of our dataset.

Firstly, the reconstructed changes in human capital are interesting in their own right. They illustrate an important aspect of global development over the past decades. Section 5 could only briefly illustrate selected trends and patterns of these remarkable increases in human capital in individual countries and on a

global scale. Much more shall and will be done in terms of systematic comparisons of national level trends which also exploit the rich detail of distributions by age and sex. Probably the single, most important lesson from this analysis of the dynamics of human capital accumulation is the great momentum and path dependence of improvements in the average educational attainment of the working-age population.

But beyond the interest in education per se, this new dataset facilitates the analysis of a great range of issues that education is assumed to influence positively. Health and survival are strongly linked with better education. Fertility levels tend to vary greatly with the level of education, and even such difficult to measure aspects of our quality of life at the societal level, the quality of institutions, the rule of law and democratic participation, are presumably facilitated by the fact that large segments of the population are well educated enough to exert the checks and balances that are necessary to establish or maintain a democracy and improve governance. For these qualities, good education of large parts of the population and not just small elites is probably a necessary, but not sufficient, condition. While such statements are currently still at the level of plausible conjectures, this new dataset will allow some real analysis and testing. In particular the study of the impacts of different distributions of human capital across categories (what mix of proportions with primary, secondary and tertiary education is most conducive to these goals under different conditions) promises to be an exciting research topic.

The greatest immediate interest in these education data clearly comes from scholars of economic growth. As mentioned above, there has been considerable concern about the fact that economic theory suggests that human capital should positively influence economic growth and at the micro level, the effects of education on individual income have been established beyond any doubt, yet the datasets available so far were not able to consistently produce significant positive effects on the macro level. Some first analyses that chose selected, well-established economic growth equations and applied them independently to both the Barro and Lee dataset as well as to our new IIASA/VID dataset, showed very promising results in the sense that the IIASA/VID data did indeed produce consistently significant positive coefficients (see Crespo and Lutz 2007). In particular the age-specific analysis seems to add to the explanatory power of the economic growth models in the sense that the growth in the human capital of younger adults (20-39) generally matters more than that of older adults, while (not surprisingly) that of pension-age men and women turns out to be irrelevant. With respect to differential impacts of the growth in populations with different levels of educational attainment, the studies indicate that for developing countries increases in younger workers with secondary education are key while in highly industrialized countries the tertiary education of more mature workers seems to matter most (Crespo and Lutz 2007).

Most economists interested in these issues have long thought that they simply have to learn to live with the highly unsatisfactory data situation and there cannot be any further improvements in the available database because they are of an historical nature and one cannot go back in time and collect new empirical data for these past periods. The fact that certain demographic methods (unknown to most economists working in this field) are now able to reconstruct such detailed historical data is a good example of the benefits of interdisciplinary collaboration and cross-fertilization.

Acknowledgement

There is a long list of people that have helped one way or another to make this work possible. Unfortunately, we cannot name them all here, but we wish to acknowledge and thank them for their contribution. A special word of thanks goes to Fernando Riosmena and Isolde Prommer for diligent data-mining and validation, to Olivier Labé and José Pessoa at UNESCO for sharing their data with us, to Patrick Gerland at the UN Population Division for valuable assistance, and to Joshua Goldstein for helping us with the methodology. We also received valuable inputs from Vegard Skirbekk, Jesus Crespo Cuaresma, Alexia Fürnkranz-Prskawetz, Sarah E. Staveteig, and Moema Figoli. Finally, thank you to Marilyn Brandl and Eryl Maedel for editing this long manuscript.

References

Alachkar, A. and W. J. Serow. 1988. "The socioeconomic determinants of mortality: An international comparison." *Genus* 44(3–4): 131–151.

Andersen, O. 1991. "Occupational impacts on mortality declines in the Nordic countries." In: W. Lutz, (ed.) *Future demographic trends in Europe and North America. What can we assume today?* London: Academic Press, pp. 41-54.

Barro, R. J. and J. W. Lee. 1993. "International comparisons of educational attainment." *Journal of Monetary Economics* 32(3): 363-394.

Barro, R. J. and J. W. Lee. 1996. "International measures of schooling years and schooling quality." *American Economic Review* 86(2): 218-223.

Barro, R. J. and J. W. Lee. 2001. "International data on educational attainment: Updates and implications." *Oxford Economic Papers* 53(3): 541-563.

Batljan, I. (Work in progress). New challenges for health human resources planning – projected changes in educational composition may result in even faster increase in number of older people than previously thought in Sweden.

Bloom, D. E. 2006. *Measuring global educational progress*. Cambridge, MA: American Academy of Arts and Sciences.

Cohen, D. and M. Soto. 2007. "Growth and human capital: Good data, good results." *Journal of Economic Growth* 12(1): 51-76.

Cohen, J., D. E. Bloom, and M. B. Malin (eds.) 2007. *Educating all Children: A global agenda*. Cambridge, MA: American Academy of Arts and Sciences.

Crespo Cuaresma, J. and W. Lutz. 2007. "Human Capital, Age Structure and Economic Growth: Evidence from a New Dataset" *IIASA Interim Report IR-07-011*. Laxenburg, Austria: International Institute for Applied Systems Analysis.

De Angelis, D., W. R. Gilks, and N. E. Day. 1998. "Bayesian projection of the acquired immune deficiency syndrome epidemic." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 47(4): 449–498.

De la Fuente, A. and R. Doménech. 2006. "Human capital in growth regressions: How much difference does data quality make? *Journal of the European Economic Association* 4: 1-36.

Doblhammer, G. 1997. *Socioeconomic differentials in Austrian Adult mortality. A study based on linked census and deaths records for the years 1981/1982.* Doctoral Thesis. Vienna: Sozial- und Wirtschaftswissenschaftliche Fakultät, Universität Wien.

Duleep, H. O. 1989. "Measuring socioeconomic mortality differentials over time." *Demography* 26(2): 345–351.

Elo, I. T. and S. H. Preston. 1996. "Educational differentials in mortality: United States, 1979–85." *Social Science and Medicine* 42(1): 47–57.

Feldman, J. J., D. M. Makuc, J. C. Kleinman, and J. Cornoni-Huntley. 1989. "National trends in educational differentials in mortality." *American Journal of Epidemiology* 129(5): 919–933.

Figoli, M. 2006. "Educational attainment in Brazil: An analysis of the rates between 1970 and 2000." *IIASA Interim Report IR-06-005.* Laxenburg, Austria: International Institute for Applied Systems Analysis.

Fotso, J.-C. 2006. "Malawi's future human capital: Is the country on track to meeting the MDGs on education?" *IIASA Interim Report IR-06-020.* Laxenburg, Austria: International Institute for Applied Systems Analysis.

Huisman, M., A. E. Kunst, O. Andersen, M. Bopp, J.-K. Borgan, C. Borrell, G. Costa, P. Deboosere, G. Desplanques, A. Donkin, S. Gadeyne, C. Minder, E. Regidor, T. Spadea, T. Valkonen, and J. P. Mackenbach. 2004. "Socioeconomic inequalities in mortality among elderly people in 11 European populations." *Journal of Epidemiology and Community Health* 58(6): 468-475.

Kitagawa, E. and P. Hauser. 1973. *Differential mortality in the United States: A study in socioeconomic epidemiology*. Cambridge, MA: Harvard University Press.

Kyriacou, G. 1991. "Level and growth effects of human capital: A cross-country study of the convergence hypothesis." *Working Paper 91-26.* New York: C.V. Starr Center for Applied Economics, Department of Economics, New York University.

Lau, L. J., D. T. Jamison, and F. F. Louat. 1991. "Education and productivity in developing countries: An aggregate production function approach. *Policy Research Working Papers, WPS 612.* Washington, D.C.: The World Bank.

Law, M., M. Lynskey, J. Ross, and W. Hall. 2001. "Back projection estimates of the number of dependent heroin users in Australia." *Addiction* 96(3): 433-443.

Lee, R. D. 1978. *Econometric studies of topics in demographic history. A dissertation in economics, Harvard University.* New York: Arno Press.

Lee, R. D. 1985. "Inverse projection and back projection: A critical appraisal, and comparative results for England, 1539 to 1871." *Population Studies* 39: 233-248.

Lutz, W. and S. Scherbov. 2004. "Probabilistic population projections for India with explicit consideration of the education-fertility link." *International Statistical Review* 72(1): 81-92.

Lutz, W., A. Goujon, and G. Doblhammer-Reiter. 1999. "Demographic dimensions in forecasting: Adding education to age and sex." In: W. Lutz, J. W. Vaupel, and D. A. Ahlburg (eds.) *Frontiers of population forecasting.* A supplement to *Population and Development Review* 24(1998): 42-58.

Nehru V., E. Swanson, and A. Dubey. 1995. "A new database on human capital stock in developing and industrial countries: Sources, methodology and results." *Journal of Development Economics* 46: 379-401.

Pamuk, E. R. 1985. "Social class inequality in mortality from 1921 to 1972 in England and Wales." *Population Studies* 39: 17-31.

Pappas, G., S. Queen, W. Hadden, and G. Fisher. 1993. "The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986." *New England Journal of Medicine* 329(2): 103–109.

Preston, S. H., M. R. Haines, and E. Pamuk. 1981. "Effects of industrialization and urbanization on mortality in developed countries." In: *International Union for the Scientific Study of Population. International Population Conference, Manila, 1981: Solicited Papers, Vol. 2.* Liege: Ordina Editions, pp. 233–54.

Sanderson, W. 2005. Education forecasts and backcasts: Lessons from the summer of 2005. Unpublished Note.

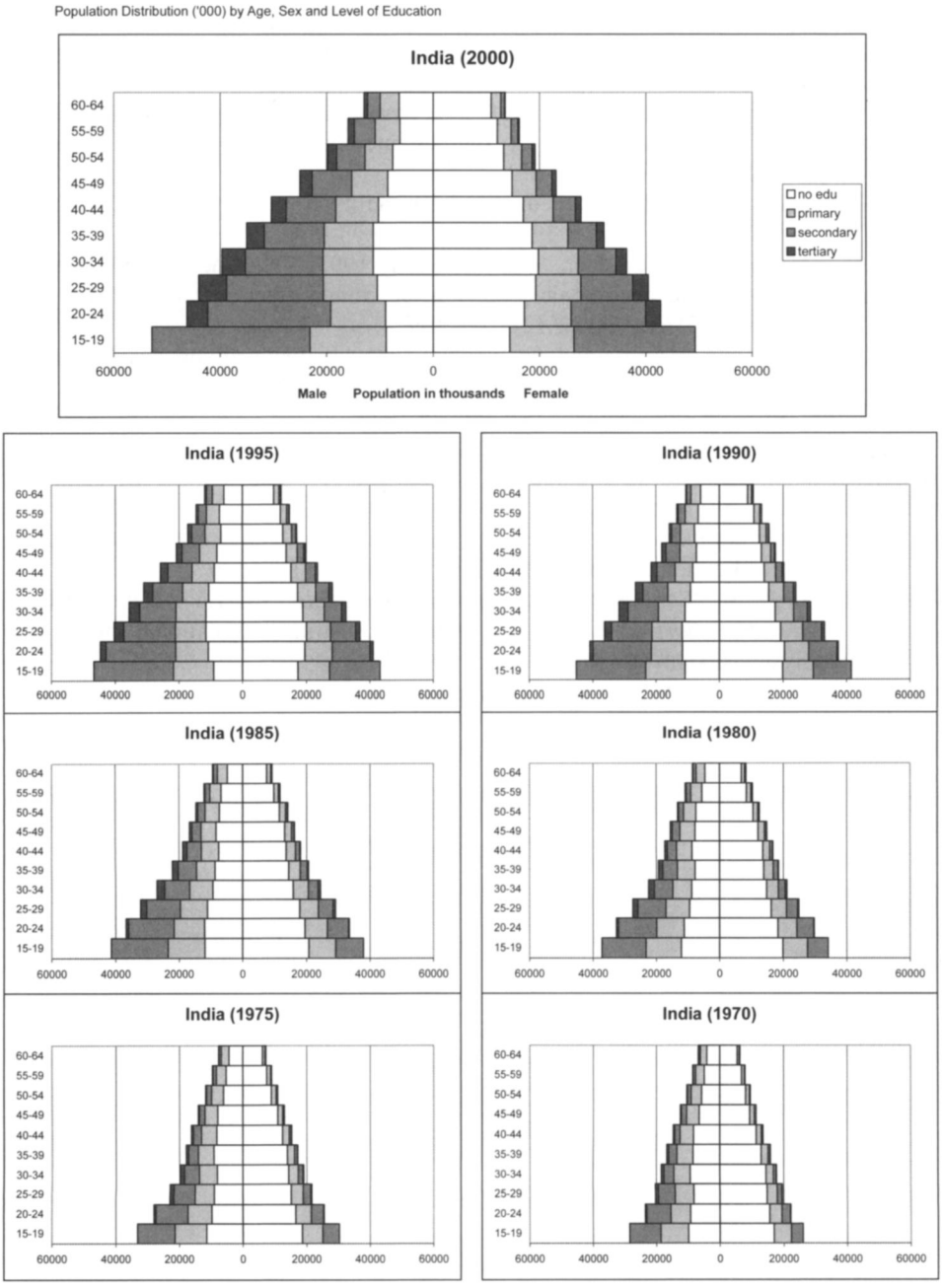
UN. 2005. *World population prospects: The 2004 revision.* New York: United Nations, Department of Economic and Social Affairs, Population Division.

Woubalem, Z. 2006. "Estimates of excess adult deaths due to HIV/AIDS in Kenya." *IIASA Interim Report IR-06-013.* Laxenburg, Austria: International Institute for Applied Systems Analysis.

Wrigley, E. A. and R. S. Schofield. 1982. *The population history of England, 1541-1871: A reconstruction.* Cambridge, MA: Harvard University Press.

Appendix A.

Standard Country Output File for the Example of India



Population Distribution ('000) by Age, Sex and Level of Education plus Means Years of Schooling

	Males					Females				
	No Edu.	Primary	Secondary	Tertiary	MYS	No Edu.	Primary	Secondary	Tertiary	MYS
2000 15-19	8867.9	14274.4	29681.6	0.0	**6.6**	14388.4	12058.4	22819.5	0.0	**5.5**
20-24	8917.9	10326.9	23088.3	3906.1	**7.2**	17138.4	8804.1	14025.1	2813.2	**5.1**
25-29	10535.7	10117.4	18153.7	5191.1	**6.9**	19288.5	8470.7	9809.7	2893.5	**4.4**
30-34	11251.9	9467.0	14566.2	4328.1	**6.4**	19787.5	7470.4	7115.4	2013.2	**3.6**
35-39	11258.2	9195.7	11243.9	3274.6	**5.7**	18600.2	6753.1	5345.5	1387.1	**3.2**
40-44	10281.5	8067.7	9242.1	2785.4	**5.6**	16974.7	5578.5	4153.2	1103.6	**2.9**
45-49	8523.5	6780.7	7453.6	2298.9	**5.5**	14878.0	4503.7	2970.6	751.0	**2.6**
50-54	7591.1	5223.6	5313.0	1683.8	**5.1**	13297.3	3362.2	1982.0	507.8	**2.1**
55-59	6255.8	4633.0	3937.7	1157.7	**4.7**	12067.1	2639.4	1209.5	262.1	**1.6**
60-64	6389.9	3614.5	2346.7	619.9	**3.6**	10929.7	1818.6	680.3	132.3	**1.2**
65+	12124.9	7331.6	3515.5	855.0	**3.2**	21509.3	3534.2	1022.4	161.2	**1.0**
15+	*101998.4*	*89032.5*	*128542.2*	*26100.7*	***5.9***	*178859.1*	*64993.3*	*71133.4*	*12025.2*	***3.5***
25+	*84212.5*	*64431.2*	*75772.3*	*22194.6*	***5.5***	*147332.3*	*44130.9*	*34288.8*	*9211.9*	***2.8***
1995 15-19	9037.9	12640.2	25056.7	0.0	**6.4**	17354.4	10005.4	15873.4	0.0	**4.5**
20-24	10732.7	10289.7	21911.3	1750.8	**6.4**	19567.9	8578.8	11851.1	971.9	**4.1**
25-29	11500.2	9656.6	16268.0	2924.7	**6.1**	20084.4	7568.3	7862.4	1352.1	**3.5**
30-34	11536.8	9399.7	11442.3	3320.6	**5.7**	18887.4	6843.0	5397.3	1396.4	**3.2**
35-39	10588.1	8281.5	9432.0	2829.1	**5.5**	17280.5	5665.8	4201.1	1112.5	**2.9**
40-44	8861.4	7020.2	7658.0	2346.7	**5.5**	15225.0	4596.4	3017.2	759.5	**2.5**
45-49	8030.3	5496.0	5533.5	1738.2	**5.0**	13759.4	3466.9	2030.7	517.2	**2.1**
50-54	6793.5	4997.1	4192.0	1218.0	**4.7**	12707.1	2765.5	1255.5	269.7	**1.6**
55-59	7240.8	4061.0	2592.5	674.2	**3.6**	11914.7	1968.1	726.2	139.4	**1.1**
60-64	5827.9	3635.1	1925.9	491.7	**3.5**	9856.3	1678.3	503.9	82.4	**1.0**
65+	10960.9	6362.3	2627.2	573.2	**2.9**	18241.4	2854.5	755.0	108.5	**0.9**
15+	*101110.5*	*81839.4*	*108639.5*	*17867.2*	***5.5***	*174878.5*	*55991.0*	*53473.7*	*6709.6*	***2.9***
25+	*81339.9*	*58909.5*	*61671.5*	*16116.4*	***5.1***	*137956.1*	*37406.7*	*25749.2*	*5737.7*	***2.4***
1990 15-19	10881.6	12495.4	21814.1	0.0	**5.9**	19852.1	9622.2	12012.6	0.0	**3.8**
20-24	11699.3	9807.2	18442.1	984.7	**5.8**	20409.8	7676.1	8857.0	454.4	**3.3**
25-29	11763.3	9564.4	12719.6	2237.6	**5.5**	19202.8	6942.6	5925.3	938.3	**3.1**
30-34	10830.8	8449.6	9579.6	2862.4	**5.5**	17578.6	5750.2	4246.3	1120.6	**2.9**
35-39	9121.9	7202.8	7810.5	2381.2	**5.4**	15531.1	4677.1	3056.3	766.4	**2.5**
40-44	8356.3	5695.0	5689.3	1774.8	**5.0**	14116.7	3546.9	2066.8	523.9	**2.1**
45-49	7208.0	5273.1	4378.2	1260.2	**4.7**	13196.2	2861.5	1290.3	275.4	**1.6**
50-54	7888.8	4394.3	2768.8	711.2	**3.5**	12607.9	2071.9	757.1	144.0	**1.1**
55-59	6654.3	4115.1	2143.9	538.5	**3.4**	10833.3	1831.2	542.0	87.5	**1.0**
60-64	5898.3	3093.4	1317.9	295.8	**2.8**	8846.3	1281.9	340.2	51.0	**0.9**
65+	9591.1	5771.8	2147.5	430.8	**2.8**	15616.8	2473.2	607.7	79.6	**0.9**
15+	*99893.6*	*75862.0*	*88811.4*	*13477.3*	***5.0***	*167791.6*	*48734.8*	*39701.5*	*4441.0*	***2.5***
25+	*77312.7*	*53559.4*	*48555.2*	*12492.5*	***4.7***	*127529.7*	*31436.5*	*18831.9*	*3986.6*	***2.0***
1985 15-19	11855.6	11670.5	17853.1	0.0	**5.4**	20725.8	8477.7	8705.4	0.0	**3.1**
20-24	11949.5	9698.4	14360.3	751.9	**5.2**	19535.7	7047.5	6622.6	315.3	**2.9**
25-29	11028.6	8585.1	10657.8	1925.2	**5.3**	17903.3	5842.7	4673.4	753.4	**2.8**
30-34	9328.2	7345.7	7926.4	2406.8	**5.4**	15830.7	4755.1	3093.1	772.6	**2.5**
35-39	8607.0	5845.7	5802.9	1800.6	**5.0**	14433.0	3616.6	2096.9	529.3	**2.1**
40-44	7516.2	5474.5	4508.2	1288.4	**4.6**	13574.8	2934.9	1315.9	279.5	**1.6**
45-49	8389.9	4647.4	2897.0	737.0	**3.5**	13142.4	2151.7	780.5	147.5	**1.1**
50-54	7290.4	4477.5	2301.3	570.9	**3.4**	11528.3	1938.4	567.9	90.8	**1.0**
55-59	6782.5	3526.7	1477.0	326.1	**2.8**	9805.3	1410.4	368.8	54.5	**0.8**
60-64	4816.4	3029.9	1286.6	267.6	**3.0**	7492.3	1264.7	344.5	48.2	**1.0**
65+	9001.9	5065.4	1550.1	280.0	**2.5**	13718.1	1996.6	416.7	47.4	**0.8**
15+	*96566.2*	*69366.8*	*70620.6*	*10354.6*	***4.6***	*157689.5*	*41436.2*	*28985.8*	*3038.3*	***2.1***
25+	*72761.2*	*47997.9*	*38407.3*	*9602.7*	***4.3***	*117428.1*	*25911.0*	*13657.8*	*2723.0*	***1.7***
1980 15-19	12124.0	11185.0	13897.8	0.0	**4.9**	19887.7	7681.7	6496.5	0.0	**2.8**
20-24	11218.9	8715.7	12077.1	647.3	**5.0**	18265.2	5946.3	5243.3	253.6	**2.7**
25-29	9515.7	7475.0	8842.2	1620.1	**5.2**	16170.3	4844.6	3397.0	520.3	**2.4**
30-34	8821.0	5972.8	5898.0	1821.9	**4.9**	14751.4	3686.1	2126.5	534.4	**2.1**
35-39	7765.3	5634.4	4608.9	1309.5	**4.6**	13916.4	3000.1	1338.0	282.8	**1.6**
40-44	8776.1	4838.1	2990.1	754.8	**3.5**	13560.5	2213.2	798.2	150.0	**1.1**
45-49	7798.0	4760.6	2420.0	594.2	**3.3**	12069.4	2021.5	587.9	93.3	**1.0**
50-54	7478.9	3860.4	1594.8	347.5	**2.7**	10492.9	1501.2	388.5	56.9	**0.8**
55-59	5600.2	3491.5	1457.4	298.1	**3.0**	8381.2	1404.1	376.9	51.9	**1.0**
60-64	4653.8	2744.6	953.4	178.4	**2.7**	6842.8	1061.3	245.0	29.7	**0.8**
65+	8412.5	4415.1	1117.4	181.7	**2.2**	11872.4	1587.0	281.4	27.8	**0.7**
15+	*92164.5*	*63093.3*	*55857.0*	*7753.6*	***4.2***	*146210.2*	*34947.2*	*21279.2*	*2000.7*	***1.8***
25+	*68821.6*	*43192.6*	*29882.1*	*7106.2*	***3.9***	*108057.3*	*21319.2*	*9539.4*	*1747.1*	***1.5***
1975 15-19	11432.0	10018.4	11758.8	0.0	**4.7**	18640.1	6470.1	5160.1	0.0	**2.5**
20-24	9734.9	7633.5	10097.2	548.3	**4.9**	16561.5	4949.8	3807.7	175.8	**2.3**
25-29	9041.1	6108.5	6625.1	1233.4	**4.8**	15135.9	3772.4	2347.1	361.6	**2.0**
30-34	7987.6	5780.0	4704.7	1331.3	**4.6**	14305.1	3075.3	1364.7	287.2	**1.6**
35-39	9077.4	4987.0	3062.6	769.0	**3.5**	13978.3	2274.9	816.0	152.6	**1.1**
40-44	8147.1	4951.9	2496.6	608.7	**3.3**	12495.1	2086.3	603.2	95.2	**1.0**
45-49	7929.5	4070.2	1663.5	359.1	**2.7**	10964.8	1562.6	401.4	58.3	**0.8**
50-54	6084.7	3767.5	1551.4	313.4	**3.0**	8912.8	1485.2	394.5	53.8	**1.0**
55-59	5247.7	3068.4	1047.8	192.9	**2.6**	7517.2	1157.2	263.2	31.5	**0.8**
60-64	4367.0	2395.8	677.5	113.5	**2.3**	6129.7	869.5	169.1	17.8	**0.7**
65+	7549.0	3685.9	772.8	113.3	**1.9**	10180.0	1245.6	186.6	15.8	**0.6**
15+	*86597.9*	*56467.0*	*44458.1*	*5583.0*	***3.9***	*134820.6*	*28948.8*	*15513.8*	*1249.6*	***1.6***
25+	*65431.1*	*38815.2*	*22602.0*	*5034.7*	***3.4***	*99619.0*	*17528.9*	*6545.9*	*1073.8*	***1.2***
1970 15-19	9945.9	8748.9	9856.3	0.0	**4.6**	16966.1	5358.5	3747.0	0.0	**2.2**
20-24	9276.6	6254.6	7593.7	418.0	**4.5**	15567.7	3868.1	2638.6	122.4	**1.9**
25-29	8211.2	5926.7	5253.1	902.4	**4.4**	14740.1	3158.6	1491.2	194.6	**1.5**
30-34	9359.4	5125.9	3129.8	782.1	**3.4**	14430.0	2340.4	834.6	155.2	**1.1**
35-39	8455.1	5119.2	2562.3	620.7	**3.3**	12937.9	2153.0	618.7	97.1	**1.0**
40-44	8309.8	4244.9	1719.0	368.0	**2.7**	11400.4	1619.0	413.2	59.7	**0.8**
45-49	6481.7	3989.3	1623.8	324.5	**2.9**	9361.2	1553.3	409.3	55.4	**1.0**
50-54	5726.2	3324.2	1119.1	203.2	**2.6**	8043.9	1231.2	277.0	32.8	**0.8**
55-59	4952.7	2693.5	748.4	123.3	**2.3**	6796.8	956.6	183.3	19.0	**0.7**
60-64	4078.6	2071.7	477.6	71.7	**2.0**	5423.5	699.3	113.8	10.2	**0.6**
65+	6672.9	3055.0	543.5	72.7	**1.7**	8797.7	996.6	128.9	9.6	**0.5**
15+	*81470.0*	*50553.8*	*34626.7*	*3886.6*	***3.5***	*124465.3*	*23934.6*	*10855.7*	*755.9*	***1.3***
25+	*62247.5*	*35550.3*	*17176.7*	*3468.6*	***3.1***	*91931.6*	*14708.1*	*4470.1*	*633.6*	***1.0***