Extended Abstract

The basis of the WIC global human capital projection: A database on population by levels of education, age and sex

Bauer, Ramon, Michaela Potančoková, Anne Goujon, and Samir K.C.

1. The framework

In 2007 and 2008, IIASA and VID published the first round of global population projections by level of educational attainment from 2000 to 2050 and the reconstruction – also called back projections – of the same data back to 1970 (Lutz *et al.* 2007[1], K.C. *et al.* 2010[2]). Both datasets required for the projections and back-projections using the cohort-component method adopted for multistate projections, base-year data on population disaggregated by levels of educational attainment by age and sex. These were gathered for 120 countries for four education categories around the year 2000. In 2011, in the framework of the Wittgenstein Centre for Demography and Global Human Capital (WIC), it was decided to implement a new round of projections for three main reasons: 1) update the previous round with more recent data from censuses and surveys; (2) increase the number of education categories from four to six to encompass more differentials in levels of attainment across the world; and 3) increase the number of countries to be able to draw a global vision of levels of educational attainment and their potential future. At first glance, the task seems simple enough: the collection of data on highest level of educational attainment that should be available from censuses, and surveys such as labour force surveys and demographic and health surveys. However, as will be shown in this paper, the task is not as trivial as one might think and suffers from many pitfalls that require extensive work to be avoided and adjustments that need to be made beyond the first requirement to obtain the data that is not always readily available.

2. Data sources

Collection of high-quality base-line data has been crucial for the human capital projections exercise. Our aim was to collect the most reliable and up to date data on population by age, sex and educational attainment for 195 countries with population of at least 100,000 in 2010. For some countries, data on educational attainment were not available at all (North Korea, Afghanistan, Papua New Guinea, Yemen) and for some others the data were at hand, but not at the sufficient level of detail or quality (Angola, Sri Lanka, Togo, Solomon Islands). In total we managed to collect and harmonize data on educational attainment by age a sex **for 170 countries** (87 % or all countries), covering 97.3 % of world population in 2010.

Our data collection focused primarily on census data which are often the best source of information on educational attainment. Register data would be ideal for our needs; however, very few countries, even among the developed ones, have population registers we could rely on. We used register data for Austria and three Nordic countries (Finland, Norway and Sweden). In most cases we had to rely on census data and if these were not reliable or of disputable quality (e.g. Nigerian census 2006) we turned our attention to household, labour force or demographic surveys.

To summarize, we have used census data for 120 countries. Our base-line data are based mostly on 2000 (94 countries) and 2010 census rounds (26 countries). For most countries 2010 round census data were not released yet at the time of the data collection process has completed, but it was possible to include 2010 censuses for a few populous countries such as Brazil, Indonesia and Japan.

[1] Lutz, W., A. Goujon, S. K.C., W. Sanderson. 2007. Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000. Vienna Yearbook of Population Research 2007, 193-235

[2] K.C., S., B. Barakat, A. Goujon, V. Skirbekk, W. Sanderson, W. Lutz. 2010. Projection of populations by level of educational attainment, age, and sex for 120 countries for 2005-2050. Demographic Research Volume 22 (15): 383 – 472.

It is evident from Figure 1 that in particularly in Africa we had to turn to survey data. When possible we used Demographic and Health Surveys (DHS). If DHS has not been carried out in the country or we had no information on educational attainment of all household members we had to look for alternative surveys, such as MICS, LSMS, PAPFAM, RHS or other household surveys. In some surveys coverage of all regions and populations can be limited and this can influence results for these countries. In particular poorest households or those located in remote areas are more likely to be omitted or refuse to participate in the survey[3]. This can lead to biased and on average more favourable educational composition of these populations compared to countries with more reliable data.

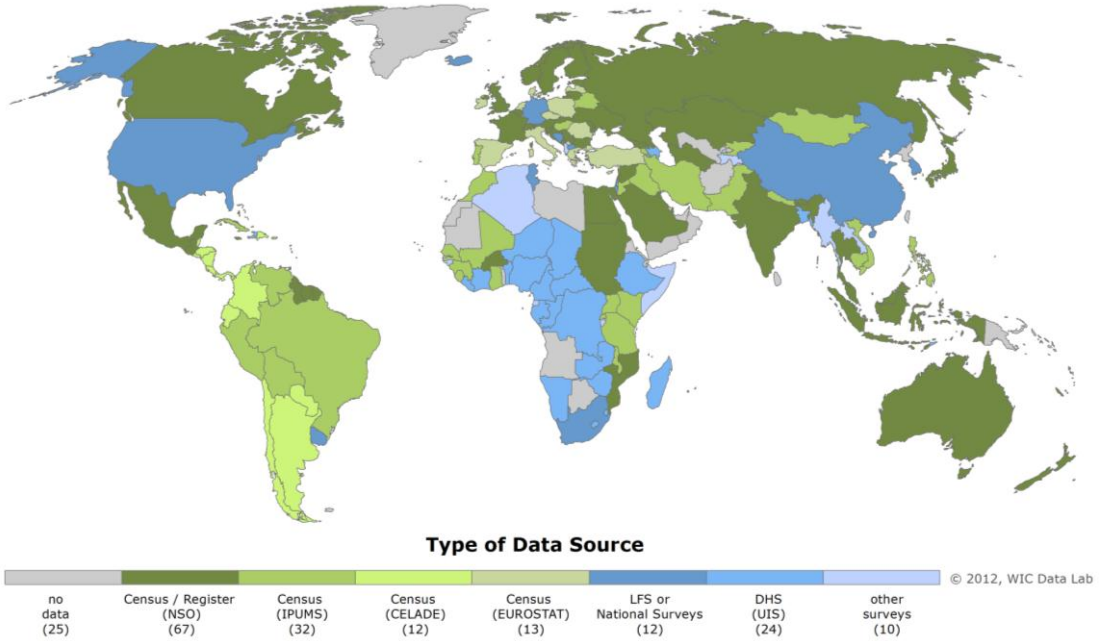

Figure 1. Data sources on educational attainment (effective June, 2012).

While for some countries it has been challenging to get any data at all, for others we collected several types of data from various sources and we further on describe our strategy of choosing the best data source.

3. Categories of educational attainment and allocation procedures

Creating our six education categories was a complex task and we have encountered several problems when allocating national-specific education categories from censuses and standardized categories in surveys to the corresponding ISCED levels. It was a challenging exercise due to several reasons:

a/ discrepancies between ISCED categories and those in the censuses or surveys, problems to identify categories to ISCED or attribute them to a single ISCED category
b/ problematic categories, such as religious education
c/ changing educating systems over time and lacking comprehensive overview of past education systems
d/ categories at post-secondary level

Table 1 provides a glimpse at the allocation rules for the educational categories.

[3] Coverage is an issue in some censuses as well, for example Sudanese census 2008 does cover only a small fraction of the population of South Sudan.

Table 1: Allocation rules for education categories

WIC 2012 categories	Allocation rules
No education	Illiterate persons; persons who have never attended school; persons who were attending 1st grade of primary education at time of survey; persons who have completed 0 years/grades at primary level (ISCED 1); persons attending adult literacy courses at time of survey; persons indicating ISCED 0 as highest educational level; khalwa (lowest level of traditional koranic schools)
Incomplete ISCED 1	grades/years of primary education below the grade of graduation from ISCED 1 is completed; persons who completed adult literacy courses; persons attending last grade of ISCED 1 at time of survey; persons who have indicated unknown number of grades/years at ISCED 1 level; traditional koranic schools above khalwa level
Completed ISCED 1	completed last grade of ISCED 1 level; completed grades below the last grade of ISCED 2 level; persons attending last grade of ISCED 2 at time of survey; persons who have indicated unknown number of grades at ISCED 2 level
Completed ISCED 2	completed last grade of ISCED 2 level; completed grades below the last grade of ISCED 3 level; persons attending last grade of ISCED 3 at time of survey; persons who have indicated unknown number of grades at ISCED 3 level
Completed ISCED 3	completed last grade of ISCED 3 level; completed number grades or years below the standard duration at ISCED 4 or ISCED 5B level; persons who have indicated unknown number of grades at ISCED 4 or 5 level
Post-secondary	Persons who have completed number of years or grades corresponding to standard duration of ISCED 4 or ISCED 5B programmes; persons with completed post-secondary university or non-university education; persons holding degrees corresponding to ISCED 4, ISCED 5B, ISCED 5A and ISCED 6 levels

4. Choosing the right dataset: Data Sources – both problem and solution

We examined various data sources with a special emphasis on detailed education and age categories for the population 15 years and older. In terms of a detailed representation of age we were targeting 5-year age groups, from age 15 onwards, at best up to age 100 years and older. With regard to detailed educational categories, we aimed to collect data at a level of detail that ensures a clear allocation to the six WIC-2012 categories. To disentangle the latent ambiguity between completed and incomplete levels of educational attainment, we collected data on both the highest level attained and highest grade/year attended whenever possible. Taking all this into account, some general rules emerged on how to identify the best – i.e. reliable, complete, detailed as well as up to date – available data on educational attainment, which further resulted in the following hierarchy of potential data sources:

1. Register or census data – based on official national population and education registers, census results (from NSOs or international statistical organisations like Eurostat or CELADE) or sample-based census micro-data (as provided by IPUMS) – usually comply with all requirements.
2. Extensive and representative sample surveys (typically LFS, National population survey for China, American Community survey for USA) – if no register data or recent census data (2000 or 2010 round) are available.
3. Household surveys on demographic and health issues (like DHS, RHS, PAPFAM) – often focus on women of reproductive age, which may not ensure a fully representative sample for men and older age groups.

4. Other household surveys (e.g. MICS, LSMS) – with more restricted samples that occasionally do not cover the entire population or territory of the country.

As a matter of fact, different data sources lead to different results. With respect to the data on educational attainment, different sources may result in different educational compositions. Generally speaking, register or census data are the best source when collecting information on the highest level of educational attainment. However, we will show many examples of inconsistencies between different survey results i.e. Eurostat vs. original data, surveys vs. census, consistency over time.

5. Adjustments and validation

The WIC-2012 dataset is based on various sources that differ by accuracy and level of detail. For that reason it was often necessary to adjust the original data in some way or other in order to obtain as detailed education and age categories as already described. In some cases interpolations were expedient to estimate 5-year age groups if the original data was not available for the corresponding age groups. Besides missing or aggregated age groups, it proved necessary to resolve the issue of categories that were either not explicitly indicated or could not been clearly allocated to one ISCED-1997 level or another. The best solution to solve such ambiguities is to allocate the categories according to the highest grade attained. If no grades were available in the original data, either additional information from other sources or analogies of populations with similar educational compositions and systems were used to split aggregated categories or to distinguish between fuzzy original categories. Due to the WIC-2012 data quality criteria, we refrained from any "guesstimation" beyond solid evidence. As a consequence, the WIC-2012 dataset includes a few countries with less than the intended six categories of educational attainment.

A special emphasis has been put on the validation of the new WIC-2012 dataset on global educational attainment. In general, whenever two or more sources were available they were validated against each other in order to reveal the more reliable source. In case that only one source was at hand that met the WIC-2012 criteria, our endeavour was to validate the data at least at a higher level of aggregation such as population 15 years and older, which is often available from official statistics. Besides case-by-case comparisons with alternative sources and before comparing the new dataset with other existing ones, it appears reasonable to validate the consistency within the WIC-2012 dataset first. In order to get a comprehensive view of the recoded data on global educational attainment, we examined the new WIC-2012 dataset by the means of maps, revealing a few discrepancies. By means of PCA, we were able to identify discrepancies across the six WIC-2012 categories within national educational compositions. The dataset was also compared against other existing datasets of educational attainment based on ISCED-1997 from UNESCO or UNSD datasets. Generally speaking, the UNESCO data tends to show higher educational levels when compared to WIC-2012, which can be explained by the WIC-2012 methodology of downgrading incomplete levels of attainment to a lower category.

6. Conclusions and outlook

Based on the work presented above, we arrived at a harmonized, consistent, and up to date dataset on levels of educational attainment for a large number of countries, which will be useful to develop research in several areas: (1) Analyse the gender, generation and geographical gaps in global educational attainment, and (2) refine the calculation of mean years of schooling as well as develop new indicators of human capital measurement and distribution. Regular updates and validation of the database will be conducted. To disseminate the database on demography and human capital to the scientific community, policy makers, and other interested individuals, we will develop and implement an online visualization tool.