ELSEVIER
www.elsevier.com/locate/worlddev

World Development Vol. xx, pp. xxx–xxx, 2016

http://dx.doi.org/10.1016/j.worlddev.2016.10.005

New Estimates on Educational Attainment Using a Continuous Approach (1970–2010)

VANESA JORDÁ and JOSÉ M. ALONSO*
University of Cantabria, Santander, Spain

Summary. — In this paper we introduce a new set of estimates on educational attainment and inequality measures of education for 142 countries over the period 1970–2010. Most of the previous attempts to measure educational attainment have treated education as a categorical variable, whose mean is computed as a weighted average of the official duration of each cycle and attainment rates, thus omitting differences in educational achievement within levels of education. This aggregation into different groups may result in a loss of information, introducing, therefore, a potential source of measurement error. We explore here a more nuanced alternative to estimate educational attainment, which considers the continuous nature of education. This "continuous approach" allows us to impose more plausible assumptions about the distribution of years of schooling within each level of education, and to take into account the right censoring of the data in the estimation, thus leading to more accurate estimates of educational attainment and education inequality. These improved estimates may help to better understand the role of education on different aspects of development.

Key words — educational attainment, inequality, mean years of schooling, measurement error

1. INTRODUCTION

The role of education in promoting human capital and well-being is an enduring issue in development studies. Although income is traditionally considered as the main indicator to measure well-being, there is a growing consensus emphasizing that the assessment of well-being should include other dimensions that may be equally relevant (Sen, 1988; Stiglitz, Sen, & Fitoussi, 2010). Among those non-income dimensions, education is acknowledged to be a key factor to measure well-being (Acemoglu & Angrist, 2001; Oreopoulos & Salvanes, 2011), and to explain health and demographic benefits (Hannum & Buchmann, 2005). At the same time, education is regarded as an important means to build up human capital, thus being a key factor in the long-run economic growth and development trajectory of nations (see, e.g., Becker, Murphy, & Tamura, 1990; Lin, 2015; Lucas, 1988; Psacharopoulos, 1994; Rebelo, 1991). Due to this key role of education in social and economic progress, it is not surprising that development scholars have devoted considerable efforts toward providing consistent estimates of educational attainment for a large number of countries over reasonably long periods of time (see, e.g., Barro & Lee, 2013; Cohen & Soto, 2007; Morrisson & Murtin, 2009).

Most of the previous attempts to measure educational attainment have focused on estimating the mean years of schooling (henceforth, MYS), since years of schooling have often been considered a good proxy for the underlying phenomenon of gaining knowledge. One of the major problems when estimating MYS is the lack of microdata on the exact number of years of schooling completed by an individual. Hence, individuals are often grouped into four broad educational levels according to their educational attainment, namely no schooling, primary, secondary, and tertiary education, which are then divided into complete and incomplete, whether the educational cycle has been finished or not. Due to this particular structure of the data, education is treated as a categorical variable whose mean is computed as a weighted average of the official duration of each cycle and attainment rates. This approach implicitly *discretizes* the variable *time that individuals attend school until leaving the educational system* to obtain an easily measurable indicator of educational outcomes. However, educational attainment is not a discrete variable but a continuous one, and the aggregation into different groups may result in a loss of information, thus introducing a potential source of measurement error.

Measurement error associated with treating education as a discrete variable may arise from different sources. First, completion rates used to break down educational levels into complete and incomplete are—in most cases—estimated. Next, the exact years of schooling are known only for those individuals classified into complete levels, corresponding to the official duration of each cycle. However, there is no information about the distribution of education within incomplete educational stages. To deal with this data limitation, the same arbitrary number of years of schooling is given to all individuals who did not complete a particular educational level. [1] This arbitrary choice may therefore bias MYS estimates and, more importantly, the direction of potential bias cannot be determined. As a consequence, the assessment of educational achievements with this kind of statistics could be questionable. Finally, it should be noted that, even if we were able to know the average years of schooling within each incomplete level, MYS estimates would be still biased. The reason is that this methodology does not take into account

* The research leading to these results has received funding from the Ministerio de Economía y Competitividad, Gobierno de España (project ECO2013-48326-C2-2-P). We are grateful to Markus Jäntti, Koen Decancq, Rolf Aaverge, Stefan Sperlich, Jose María Sarabia, Miguel Niño Zarazúa, Charlie Ovink, and participants of the X Winter School on Inequality and Social Welfare Theory, the Sixth ECINEQ meeting and the CSB Seminar at the University of Antwerp for their valuable comments. We also thank the Editor and two anonymous referees for their helpful comments and suggestions. Final revision accepted: October 8, 2016.

Please cite this article in press as: Jordá, V., & Alonso, J. M. New Estimates on Educational Attainment Using a Continuous Approach (1970–2010), *World Development* (2016), http://dx.doi.org/10.1016/j.worlddev.2016.10.005

the right censoring of census data. All individuals classified as having completed tertiary education are supposed to have finished university studies, but it is conceivable that some of them may have been enrolled in masters or Ph.D programs. Therefore, assuming that individuals within this group received the same years of schooling may bias downward MYS estimates.

In this paper, we explore a more nuanced alternative, which considers the continuous nature of education. This "continuous approach" allows us to impose more plausible assumptions about the distribution of years of schooling within each level of education, and to take into account the right censoring of the data in the estimation. We argue that a flexible parametric specification must be considered. More specifically, we employ the generalized gamma (GG) distribution to model the time that individuals attend school until they complete the educational cycle or decide to drop out.

Years of schooling have not only been used to assess educational attainment, but also to analyze the evolution of international inequality in education (see Ram, 1990; World Bank, 2005) and to approximate the national distribution of schooling (Castelló-Climent, 2008; Meschi & Scervini, 2013; Thomas, Wang, & Fan, 2001). Nonetheless, since estimates of the global and national distributions of schooling are usually derived from estimates of MYS, it is conceivable that they may also suffer from measurement error. [2] Since education inequality has been often considered as a key factor explaining different aspects of major interest among development scholars and practitioners, such as poverty, life expectancy, and income inequality (see, e.g. Castelló-Climent & Doménech, 2008; Gregorio & Lee, 2002), we also present a set of educational inequality measures derived from our MYS estimates, which should be helpful for future empirical work.

The main contribution of this paper resides, therefore, in the development of a new set of estimates on educational attainment and inequality measures of educational outcomes for 142 countries from 1970–2010, taking into account the distribution of years of schooling within each level of education. The full set of estimates is available at *www.educationdata.unican.es*. In the following section, we briefly explore previous attempts to measure educational attainment, before describing our methodology to get more reliable estimates of MYS and measures of educational inequality. Thereafter, we introduce the main features of our new estimates, and compare them with those provided by Barro and Lee (2013) and Benaabdelaali, Hanchane, and Kamal (2012). We conclude the paper by considering the practical implications of our study.

2. A NEW APPROACH TO ESTIMATING EDUCATIONAL ATTAINMENT

To estimate educational attainment, two types of information are required; firstly, data on educational attainment of each level of education and, secondly, information about their official duration in each country to transform the categorical variable into a cardinal indicator. The first type of data can be drawn from census and surveys compiled by different national and international statistical agencies. Individuals are classified into four different categories: no formal schooling, primary, secondary, and tertiary schooling depending on the last year of education attained. When individual data are not available, the proportion of the population in each level of education is extrapolated by making specific assumptions regarding mortality rates and migration trends.

(a) *Previous measurements of educational attainment*

The development of homogeneous educational attainment estimates for a large number of countries has attracted significant attention from scholars over the last three decades (De la Fuente & Domenech, 2013). Several data sets on educational attainment are available now, most of them covering the period after 1960. [3] (Barro & Lee, 2013; Cohen & Soto, 2007; De la Fuente & Doménech, 2006; Lutz, Goujon, KC, & Sanderson, 2007; Psacharopoulos & Arriagada, 1986) Among these existing data sets, Barro and Lee (2013) (henceforth, BL) is the most comprehensive database in terms of time and geographical coverage, including information for 146 countries at five-year intervals over the period 1950–2010.

The latest version of the BL database improves their previous estimates on MYS, incorporating most of the criticisms pointed out in the literature (see Cohen & Soto, 2007; De la Fuente & Doménech, 2006). One of the main drawbacks of earlier versions of the BL series was the need to consider constant mortality rates across age groups, since aged individuals are expected to have higher mortality rates. The new version of the BL series overcomes this potential shortcoming by taking into account the heterogeneity in mortality rates across age groups. An additional improvement is the adjustment of mortality patterns by educational level, which allows one to consider the possibility that more educated people would have a lower probability of dying. The extrapolation method has been also modified to adopt the methodology proposed by Cohen and Soto (2007), which seems to have led to more consistent estimates. Particularly, attainment rates for the age cohort k in $t-5$ are calculated from census data of the age cohort $k+5$ in the year t, using the mortality patterns of the country.

BL data on educational attainment classify individuals into seven categories: no schooling, incomplete primary, complete primary, incomplete secondary, complete secondary, incomplete tertiary and complete tertiary. At a first stage, educational attainment rates are calculated for the four general categories using census data. Thereafter, these proportions are broken down into complete and incomplete categories using data on completion rates. Completion rates also come from census data, so an extrapolation method is used to estimate missing observations, which may be another potential source of bias.

MYS are then computed using attainment rates and the official durations of educational cycles in each country. [4] Formally, for country i in time t,

$$mys_{i,t} = \sum h_{i,t}^{(j)} D_{i,t}^{(j)}$$

where $h^{(j)}$ is the attainment rate of the educational level $j = (noschooling, primary, secondary, tertiary)$ $(complete, incomplete)$, and $D^{(j)}$ indicates the official duration of the education cycle in years.

To illustrate this, let us assume that the distribution of education in country i at time t follows the pattern depicted in Table 1. The simplest approach would be to consider only the complete information in our example, i.e., without disaggregating attainment rates in complete and incomplete, giving therefore the official duration of the cycle to all individuals classified in each level. MYS would be then given by $mys = (0.0025 * 0) + (0.3213 * 6) + (0.4609 * 12) + (0.2153 * 16)$, which clearly overestimates the mean. The strategy of breaking educational levels into complete and incomplete would potentially solve this problem, but there are no data about the exact years that a person has been

Please cite this article in press as: Jordá, V., & Alonso, J. M. New Estimates on Educational Attainment Using a Continuous Approach (1970–2010), *World Development* (2016), http://dx.doi.org/10.1016/j.worlddev.2016.10.005