New estimates on educational attainment using a continuous approach (1970-2010)

Vanesa Jorda and Jose M. Alonso

Department of Economics, University of Cantabria

forthcoming in *World Development*

Abstract

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 to 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.

Keywords: 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 et al., 2010). Among those non-income dimensions, education is acknowledged to be a key factor to measure well-being (Acemoglu and Angrist, 2001; Oreopoulos and Salvanes, 2011), and to explain health and demographic benefits (Hannum and 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., Lucas, 1988; Becker et al., 1990; Rebelo, 1991; Psacharopoulos, 1994; Lin, 2015). Due to this key role of education in social and economic progress, it is not surprising that development scholars have devoted considerable efforts towards providing consistent estimates of educational attainment for a large number of countries over reasonably long periods of time (see, e.g., Cohen and Soto, 2007; Morrisson and Murtin, 2009; Barro and Lee, 2013).

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 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 downwards 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 (Thomas et al., 2001; Castello and Domenech, 2008; Meschi and Scervini, 2013). 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

---

[1]For instance, MYS provided by UNESCO are calculated assuming that people who have not finished the educational cycle have successfully completed half of the official duration of such a level. Then, these estimates would be always biased except in the hypothetical case that population distributes uniformly across years within each level of education, which would be a coincidence rather than the norm.

[2]Results of studies dealing with international inequalities in education are expected to represent a lower bound of inequality given that inequality within countries is not being considered. This kind of analysis

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. DeGregorio and Lee, 2002; Castello and Domenech, 2008), 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 to 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 et al. (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.

---

would provide, *a priori*, valuable information in terms of disparities. However, since MYS are biased indicators and due to the fact that the bias is not systematic, these results are no longer lower bounds.

2.1 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 and Domenech, 2013). Several data sets on educational attainment are available now, most of them covering the period after 1960.[3] (Psacharopoulos and Arriagada, 1986; De la Fuente and Domenech, 2006; Cohen and Soto 2007; Lutz et al. 2007; Barro and Lee, 2013) Among these existing data sets, Barro and Lee (henceforth, BL) (2013) 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 De la Fuente and Domenech, 2006; Cohen and Soto, 2007). 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.

---

[3]With the exception of Morrisson and Murtin (2009; 2013), who computed estimates of the proportions of the population that attained the four broad educational levels for a very long period that covered the last two centuries. However, this data set includes information for 32 macro-regions rather than at country level.

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,

$$mys_{i,t} = \sum h_{i,t}^{(j)} D_{i,t}^{(j)} \qquad (1)$$

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 is no data about the exact years that a person has been attending a particular educational level, hence an arbitrary value is set for individuals pertaining to this group. For instance, MYS computed by UNESCO are obtained assuming that people who have not finished a cycle have attained half of its official duration. This assumption can yield under or overestimated figures on MYS depending on the distribution of population within each educational level and, more importantly, it is not feasible to figure out the direction of the bias.

Table 1: Attainment rates and official duration of different educational levels

Educational level	Attainment rate	Official duration
No schooling	0.0025	0
Primary incomplete	0.2413	unknown
Primary complete	0.0800	6
Secondary incomplete	0.2973	unknown
Secondary complete	0.1636	12
Tertiary incomplete	0.1248	unknown
Tertiary complete	0.0905	16

Educational attainment rates of each educational level represent points of the cumula-

---

[4]UNESCO's Statistical Yearbook reports country-specific duration figures for each educational level

tive distribution function (CDF) of the variable length of schooling. In this sense, we can consider the variable years of schooling as a *discretization* of the continuous variable *time that individuals attend school until leaving the educational system*, which is used to obtain an easily measurable indicator of educational outcomes. Most of the previous studies estimate mean levels of education by linking these points linearly, which implicitly assumes that the population distributes uniformly across academic years. Hence, the mean of that part of the distribution would be given by half of the years of the complete level.

A graphical example may help to clarify this point; squares in Figure 1 represent the attainment rates of complete educational levels, for which we know the number of years of schooling associated with the official duration of the cycle. In contrast, this information is not known for incomplete levels. These values can range from the maximum years of schooling of the previous educational level up to the maximum years of its own cycle, represented in the graph by the black lines at each incomplete level. The dashed line represents the linear interpolation applied in most previous studies. The level of education assigned to the incomplete levels assuming this kind of distribution is given by half of the official duration (triangles placed at the middle of each line). The distribution of time of schooling, which is indeed continuous, would be represented by the solid line. This simple example reveals that it is not possible to find out the direction of the bias. If our choice of the number of years schooling given to the incomplete level lie above (below) the actual mean of the incomplete level, the MYS of the educational cycle would be overestimated (underestimated). The positive or negative bias on the overall MYS is given by the sum of the bias for all levels of education.

It is worth highlighting that even though we were able to know the exact values associated with incomplete attainment rates, our estimates could be still biased, since discrete approaches do not take into account the right censoring of census data. For instance, individuals who have completed a master's degree or higher are considered as having the same number of years of schooling as males or females who have completed just a bachelor's degree or the like, an assumption which may bias downwards MYS estimates.

Following from this, previous attempts to estimate educational inequality based on MYS estimates would suffer from the same sources of bias as the mean (see, for example, two recent studies by Castello, 2010, and Benaabdelaali et al., 2012). In addition, although splitting each educational level in two groups clearly increases the number of available points of the empirical distribution of schooling, we would still be working with grouped
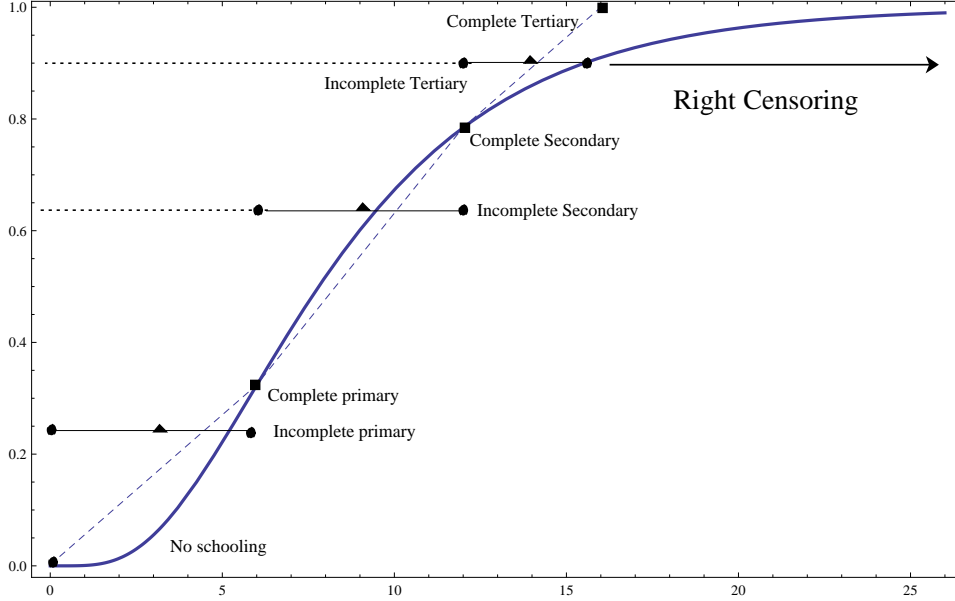
Figure 1: Sources of bias of mean years of schooling

data. Hence, computing empirical inequality measures with this kind of limited information may lead to biased results, because the variability within each educational cycle is not considered.

2.2   Methodology

In this paper, we use a parametric model to provide a more reliable approximation of the shape of the distribution between the empirical points of the CDF. Let X be a continuous variable representing the time of schooling until either completing the maximum level of education or dropping out of school. Let $h_i^{(j)}, j = NS, P, S, TI, TC$ be the share of population that attained the educational cycle j (no schooling (NS), primary (P), secondary (S), incomplete tertiary (TI) and complete tertiary (TC)) in country i; and denote by $D_i^{(j)}$ its official duration. Educational attainment rates for the first three broad categories represent the frequencies associated with population that attended school during the number of years of official duration or less, $h_i^{(j)} = Pr[X \le D_i^{(j)}]$. As regards the last category, we need to decompose attainment rates of tertiary education to take into account the right censoring of the variable. To do so, we assume that $H_i^{(TI)}$ informs about the proportion of population that attained $D_i^{(T)}$ number of years of schooling or less, and $H_i^{(TC)}$ represents the proportion of population that attained $D_i^{(T)}$ years of schooling or more. Therefore, we only use

data on official durations without the need to assign arbitrary values for incomplete levels.

We use BL data on educational attainment for the four broad educational levels, without considering the decomposition of schooling rates into complete and incomplete, except for the highest category, i.e. tertiary education. Data on the official duration of primary and secondary education are drawn from UNESCO (2014), which includes series from 1970 onwards. We take into account changes in the educational system over time, assuming that those changes apply to the whole population. Following UNESCO's definition, illiterate people are considered to have less than one year of education.[5] Finally, following previous studies such as BL, Cohen and Soto (2007), and UNESCO (2013), we assume that the duration of tertiary education is equal to four years in all countries over the whole period.[6]

After describing the data used to estimate educational attainment, we need to define the modeling strategy. Survival analysis provides a suitable framework to model the time that individuals attend the school until leaving the educational system. This methodology has been extensively used to model the probability of dropping out of school (see, for example, Arulampalam et al., 2004; Plank et al., 2005). This is the first study, to our knowledge, applying survival techniques to model the overall distribution of education.

The main assumption of our methodology is that the same functional form is used to model the distribution of years of schooling in all countries over the whole period. Empirical evidence, however, suggest that the distribution of education changes both over time and across countries (Morrisson and Murtin, 2009). For example, one mode would be expected in developed countries with compulsory years of schooling, while zero mode distributions seem characteristic of developing countries, exhibiting high illiteracy rates. To deal with these potential changes in the distribution of education, we have to deploy a flexible model whose parameters can vary across countries and over time, thus allowing us to model particular features of the distribution of schooling.

There are several alternatives to accommodate time and country varying parameters, from the simplest model given by the exponential distribution to the non-parametric and semiparametric families of life distributions. Among the whole range of potential alternatives, the generalized gamma (GG) distribution seems to be the most appealing option.

---

[5]For a comprehensive guide about how to use and interpret UNESCO's census data see http://www.uis.unesco.org/Library/Documents/hhsguide04-en.pdf.

[6]The choice of four years of schooling is due to the high heterogeneity in the duration of tertiary education programs. On average, duration of completed short-cycle tertiary education (ISCED 5) was 3.7 between 2000 and 2010 (see UNESCO, 2013).

The reason is twofold; first, the GG distribution is a parsimonious model that nests most of the parametric assumptions described in the literature (see, e.g., Marshall and Olkin, 2007). Second, the GG distribution is able to model one and zero mode distributions and to represent several types of hazard rates, which, in our case, can be defined as the probability of dropping out of school at time t, given that the individual attended the school up to time t. This distribution is, therefore, closely linked to dropout patterns, which are expected to differ substantially across countries (Filmer and Pritchett 1999; Lewin 2009). Hence, the flexibility of the GG distribution to consider such heterogeneity, makes it an outstanding candidate to model the distribution of education. It is important to highlight that this parametric model includes as particular cases most of the distributions commonly used in survival analysis, including the Weibull, the exponential and the gamma distributions. Hence, the GG distribution would converge to any of its special cases if needed.

The GG distribution can be defined in terms of the probability density function (PDF) as follows (Stacy, 1962):

$$f_{GG}(x; a, \beta, p) = \frac{a x^{ap-1} e^{-(x/\beta)^a}}{\beta^{ap} \Gamma(a)}, \quad x \geq 0, \tag{2}$$

where $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$ is the gamma function, $\beta > 0$ is a scale parameter, $a > 0$ and $p > 0$ are shape parameters.

The CDF is given by the following expression:

$$F(x; a, \beta, p) = IG(x^a; p, \beta^a) = \frac{1}{\Gamma(p)} \int_0^{(x/\beta)^a} t^{p-1} e^{-t} dt$$

where $IG(.)$ denotes the incomplete gamma function.

The distribution of education of each country is estimated by non-linear least squares. We use information on the share of the illiterate population and attainment rates of primary, secondary, and incomplete and complete tertiary education. The strategy is to minimize the sum of squared deviations between the empirical attainment rates $(h^{(j)})$ and the associated theoretical probabilities for each level of education $(Pr[X \leq D^{(j)}])$, modeled by the CDF of the GG distribution. The right censoring of education data is taken into account by modeling the probability associated to the complete tertiary category using the survival function $(1 - F(x))$ instead of the CDF used for the first educational four levels.

For each country $(i = 1, ..., 142)$ and year $(t = 1970, ..., 2010)$, the function to be

minimized is given by,

$$\min_{a,\beta,p} \sum_{j=1}^{J-1} \left(IG(v_{it}^{(j)}; p, \beta^a) - h_{it}^{(j)} \right)^2 + \left(1 - IG(v_{it}^{(TC)}; p, \beta^a) - h_{it}^{(TC)} \right)^2, \qquad (3)$$

where $v_{it} = \left(D_{it}^{(j)} \right)^a$.

Non-linear regression techniques involve the definition of starting values for the optimization algorithm. The estimation of the GG distribution is characterized by multiple local minima. This fact makes it difficult to ensure that the estimated values are those which globally minimize the residual sum of squares (RSS). To assess the robustness of our estimates, we consider a set of alternative initial values, using an iterative method inspired in the algorithm proposed by Gomes et al. (2008). Let us consider the following relationship between the gamma and the GG distributions,

$$X \sim GG(a, \beta, p),$$

then

$$X^a \sim G(p, \beta^a).$$

We define a grid for parameter a from 0.2 to 20 by steps of 0.2, and compute the moment estimates for the parameters of the gamma distribution of the variable X^a. These estimates, along with the grid value, are used as starting values for the optimization algorithm[7]. This process is repeated for all values in the grid, using the estimation that reports the lowest RSS.

Once the parameters are estimated, MYS are obtained by substituting them in the expression of the mean of the GG distribution:

$$\mu = \frac{\beta \Gamma(p + \frac{1}{a})}{\Gamma(p)}. \qquad (4)$$

In addition to MYS estimates, we present several education inequality measures which satisfy the Lorenz ordering. The Lorenz curve is a powerful tool to compare and order distributions according to their inequality levels. If these curves do not cross, the closest distribution to the egalitarian line would be declared as less unequal by any inequality

---

[7]We use the package *optim* in *R* to find the minimum of Eq.(3). BFGS algorithm is implemented by default and L-BFGS is used when this method reports error. The computation of the gradient is done numerically.

measure consistent with the Lorenz ordering.

Following Sarabia and Jorda (2014), the Lorenz curve can be generally expressed as,

$$L(u) = F_{X_{(1)}}(F_X^{-1}(u)), \tag{5}$$

where $F_Y^{-1}(u)$ denotes the quantile function and $F_{Y_{(1)}}(x) = \int_0^x t f(t)dt$ is the distribution of the first incomplete moment, defined for the GG distribution as follows:

$$F_{X_{(1)}}(x; a, \beta, p) = IG(x^a; p + 1/a, \beta^a)$$

Substituting the explicit expressions of these distributions for the GG family in Eq.(5), we obtain the Lorenz curve,

$$L(u; a, p, \beta) = IG\left[\left(IG^{-1}(u; p, \beta^a)\right)^a; p + \frac{1}{a}, \beta^a\right],$$

where $u \in [0, 1]$ and $IG^{-1}(.)$ stands for the inverse of the incomplete gamma function.

However, the Lorenz ordering is partial in the sense that not all distributions can be ranked. In these cases, we need to use inequality measures that provide a complete ordering of distributions. The Gini index has been the main indicator used to measure income inequality and its popularity has spread to education variables, whose disparities have been mainly measured using this indicator (Thomas et al., 2001). This index is sensitive to the middle of the distribution, and it does not allow us to change the weight given to differences in specific parts of the distribution. Varying the sensitivity of inequality measures to the bottom or the upper tails is particularly relevant when there is no Lorenz dominance. If two Lorenz curves cross, inequality measures can point out different results depending on their sensitivity to different parts of the distribution. For this reason, we compute an alternative set of inequality measures belonging to the GE family, which includes a sensitivity parameter for setting the importance given to the differences at the upper tail.

To examine the evolution of educational inequality, we compute the GE measures for different parameter values. The mean log deviation (MLD) corresponds to the GE index when the parameter is set to 0, thus being more sensitive to the bottom part of the distribution. The case given by the Theil's entropy measure is equally sensitive to all parts of the distribution, being characterized by a parameter value equal to 1. We also compute the GE measure when the sensitivity parameter is set equal to 2, which gives more weight

to differences in higher education.

For the GG distribution, GE measures are given by the following expressions (Jenkins, 2009; Sarabia et al., 2016):

$$GE(2) = -\frac{1}{2} + \frac{\Gamma(p + \frac{2}{a})\Gamma(p)}{2\Gamma^2(p + \frac{1}{a})},$$

$$T_0 = -\frac{\psi(p)}{a} + \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)}, \tag{6}$$

$$T_1 = \frac{\psi(p + \frac{1}{a})}{a} - \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)}. \tag{7}$$

where $\psi(z) = \Gamma'(z)/\Gamma(z)$ denotes the digamma function.

This family of GE measures are additively decomposable in two components, that can be obtained using information on the mean, population shares and national inequality measures. The consideration of additively decomposable measures makes these data extremely useful, since it allows for the computation of overall inequality for any combination of countries. Such flexibility would not be achieved if we had focused solely on the Gini index, which involves, in addition to the between-country and the within-country components, an overlapping term that is specific for the particular group of countries under consideration.

The distribution of education of a particular group or region of countries is defined as a mixture of the national distributions, weighted by their population shares. Let $X_i, i = 1, ..., N$ be the length of schooling in the county i which is assumed to follow a GG distribution given by Eq.(2). Then, the regional PDF of education can be expressed as,

$$f(x) = \sum_{i=1}^{N} \lambda_i f_i(x), \tag{8}$$

where λ_i stands for the population weights of the countries. The regional CDF would be the integral of the PDF given by Eq.(8),

$$F(x) = \sum_{i=1}^{N} \lambda_i F_i(x) = \sum_{i=1}^{N} \lambda_i IG(x^a; p_i, \beta_i^{a_i}),$$

The mean of this distribution would then yield the regional MYS,

$$\mu = \sum_{i=1}^{N} \lambda_i \mu_i = \sum_{i=1}^{N} \lambda_i \frac{\beta_i \Gamma(p_i + \frac{1}{a_i})}{\Gamma(p_i)}.$$

Then, the mean of a particular group of countries is given by a weighted average of the national means and the population shares.

The regional estimates of the GE measures can be easily derived by taking advantage of the decomposition of this family. The general expression of the GE measure is given by,

$$I(\theta) = \sum_{i=1}^{N} \lambda_i^{1-\theta} s_i^{\theta} I_i^{(\theta)} + \frac{1}{\theta(\theta-1)} \left(\sum_{i=1}^{N} \lambda_i \left(\frac{\mu_i}{\mu} \right)^{\theta} - 1 \right), \qquad (9)$$

where λ_i and $I_i^{(\theta)}$ are the population share and the GE measure of the country i. s_i stands for the proportion of mean income of the country i in the overall mean of the group: $s_i = \frac{\lambda_i \mu_i}{\mu} = \frac{\lambda_i \mu_i}{\sum_{i=1}^{N} \lambda_i \mu_i}$.

The especial cases given by the Theil and the MLD can be decomposed as follows:

$$T_W = \sum_{i=1}^{N} s_i T_i; T_B = \sum_{i=1}^{N} s_i log \left(\frac{\mu_i}{\mu} \right),$$

$$L_W = \sum_{i=1}^{N} \lambda_i L_i; L_B = \sum_{i=1}^{N} \lambda_i log \left(\frac{\mu}{\mu_i} \right),$$

where T_i and L_i are, respectively, the Theil (Eq.(7)) and the MDL (Eq.(6)) indices for country i.

2.3 Goodness of fit

As already mentioned, our estimates rely on the assumption that schooling follows a GG distribution. Therefore, before moving onto the analysis of the evolution of educational patterns, we should investigate the goodness of fit (GOF) of the model. It should be noted that, because we have used grouped data, we do not have information about the size of the sample, thus conventional tests of GOF cannot be applied, because the distribution of the statistic is unknown.[8] As a measure of GOF, we have computed the RSS for each country/year distribution.[9] This measure illustrates the size of the differences between the observed attainment rates and the estimated ones using the GG distribution. The smaller the RSS, the closer the values of the estimated and the observed educational achievements.

We focus, therefore, on the performance of national estimates using the RSS for both,

---

[8]Bootstrapping techniques cannot be used in a context of limited information because we do not know the size of the sample, which has a strong impact on the bootstrap p-values.

[9]Parameter estimates and residual sum of squares for all distributions are available upon request.

Table 2: Goodness of fit: quartiles of residual sum of squares by year

	year	1st Cuartile	Median	3th Cuartile
Total	1970	1.58E-05	1.99E-04	1.07E-03
	1980	4.65E-05	6.57E-04	2.07E-03
	1990	2.65E-04	1.10E-03	3.32E-03
	2000	2.30E-04	1.97E-03	5.37E-03
	2010	7.35E-04	2.86E-03	6.88E-03
Males	1970	4.81E-05	4.50E-04	1.60E-03
	1980	1.28E-04	9.06E-04	3.24E-03
	1990	4.79E-04	1.75E-03	4.37E-03
	2000	4.63E-04	2.64E-03	6.70E-03
	2010	6.94E-04	3.25E-03	8.53E-03
Females	1970	3.77E-06	7.26E-05	5.06E-04
	1980	2.57E-05	2.45E-04	1.56E-03
	1990	1.14E-04	7.42E-04	2.43E-03
	2000	2.32E-04	1.21E-03	4.86E-03
	2010	5.14E-04	2.62E-03	5.94E-03

the entire population and disaggregated by gender. A summary of this information is shown in Table 2, which includes the three quartiles of the distribution of this statistic. The first quartile comprises the best 25 percent of the fits, the median the best 50 percent and the third quartile the 75 percent. Overall, RSS values are reasonably good. The mean deviations of the estimated attainment rates and the observed ones are at most 4 percent (resulting in a RSS value of 0.008) in the 75 percent of the estimations performed. These preliminary results suggest that the novel approach proposed in this paper fits the observed attainment rates accurately and, hence, yields reliable estimates of MYS and educational inequality measures.

3   Data overview

The new estimates on educational attainment include data for the 142 countries over the period 1970 to 2010. These estimates include information for population aged over-25 and over-15 for both, total population and disaggregated by gender. In this section, we provide an overview of these new data figures, starting with the estimation of MYS, before presenting estimates on the proposed educational inequality measures. Table 3 summarizes the main trends related to MYS of the population aged 15 years and over, by income level and gender. Our estimates suggest, on average, that individuals across the globe are

completing more years of schooling with every passing decade. Although there still seem to be some differences across countries and regions, one could observe substantial increases in educational attainment, particularly in low and middle income-countries and, particularly, for females.

In 1970, among individuals aged 15 and older in high-income countries, males received, on average, 6.35 years of education, and females acquired 5.49 years; by 2010 the average increased to 11.18 years for men and 10.88 years for women. These estimates not only reflect an upward trend in educational attainment, but also that female years of schooling have increased relative to that of males during the last decades, thus leading to gender convergence in terms of educational attainment in high-income countries. Regarding low and middle-income countries, the average number of years of education is much lower than in high income countries, but the rate of increase is noticeably higher (about 171.12 percent *versus* 87.51 percent). On average, years of education rose from 2.70 years for males and 1.75 years for women in 1970, to 6.70 years and 5.51 years in 2010. These estimates indicate that, although there is still a significant gender gap in this group of countries, females are getting closer to males in terms of educational attainment, which is contributing to closing that gender gap in low and middle-income countries too.

It should be noted that, although years of schooling are widely used to measure the educational performance of a country and to make international comparisons, this indicator only considers the mean of the distribution of educational outcomes. Then, even when countries progress on average, it does not imply that the whole society improves its educational levels. To provide a complete picture of the evolution of global educational achievements, Figure 2 presents the global CDF of educational outcomes for each decade from 1970 to 2010, which is computed as a mixture of the national distributions (Eq.(2.2)). This is the first study, to our knowledge, showing the global distribution of schooling instead of limiting the analysis to the frequencies associated to the four levels of education. The most relevant fact is the existence of first order dominance of all years over the preceding ones. This means that the progress in education has been achieved at all levels, a pattern that is rather contrasting with the evolution of income distribution (see Chotickapanich et al., 2012; Jorda et al., 2014).

To provide a clearer picture about the evolution of each part of the distribution, Figure 3 shows the global PDF of education. Our results characterize education by a zero mode distribution during almost the whole period. This shape is driven by the high pro-

Table 3: Mean years of schooling (population 15 age and over)

Year	All countries (142)			High income (52)			Low and middle income (90)		
	Total	Male	Female	Total	Male	Female	Total	Male	Female
1970	3.4405	3.8599	3.0399	5.8550	6.3463	5.4927	2.2461	2.7032	1.7526
1975	3.8813	4.3770	3.4125	6.5221	7.1263	6.0885	2.6433	3.1607	2.0853
1980	4.3948	4.9867	3.8614	7.1169	7.7468	6.6863	3.2011	3.8414	2.5542
1985	4.8346	5.4101	4.3004	7.7464	8.4533	7.3231	3.6583	4.2436	3.0153
1990	5.2144	5.6868	4.7292	8.4665	8.9848	8.0342	3.9905	4.5040	3.4254
1995	5.7310	6.2916	5.1944	9.1075	9.5173	8.7763	4.5311	5.1943	3.8660
2000	6.2274	6.7482	5.7644	9.7483	10.0850	9.4740	5.0455	5.6734	4.4682
2005	6.7449	7.3003	6.2449	10.3871	10.6692	10.2262	5.5897	6.2722	4.9339
2010	7.2189	7.6964	6.7904	10.9786	11.1775	10.8810	6.0898	6.6890	5.5168
Δ1970-2010 (%)	109.82	99.39	123.38	87.51	76.13	98.10	171.12	147.45	214.78

Note: Data computed for the beginning of each five-year period.

portion of illiterate population, which is particularly relevant in developing countries of Saharan-Africa and Asia (United Nations, 2014). Within the last decade, the shape of the distribution changed completely. A new mode emerged around six years of schooling while the proportion of illiterate individuals decreased substantially. In fact, the estimated proportion of people with less than one year of schooling fell from 35 percent to 12 percent over the last 40 years, which involves a reduction of 66 percent. Interestingly, the proportion of people that attained primary education decreased from 40 to 27 percent of the world population. The rationale behind this finding is that the probability mass moved away from the illiterate category to primary schooling and, at the same time, from primary to secondary schooling. Indeed, secondary schooling rates substantially increased over the last decades, due to the convergence process in compulsory years of schooling (Murtin and Viarengo, 2011) and the promotion of education in East Asia (Baker and Holsinger, 1996). Finally, tertiary education was characterized by an outstanding increase from 4 percent in 1970 to 16 percent in 2010. The improvement of higher education is mainly driven by developed countries, although developing countries in Asia and Latin America seem to have played also a key role in the last years.

Turning now our attention to the proposed inequality measures, we first investigate the evolution of disparities graphically, before quantifying the global levels of inequality in education. Figure 4 shows the Lorenz curves for years 1970, 1990 and 2010, which reveal a Lorenz dominance pattern over the period under analysis. This result implies that any measure consistent with the Lorenz ordering would reveal a reduction of inequality levels in educational outcomes. Since GE measures satisfy the Lorenz ordering, all these indices
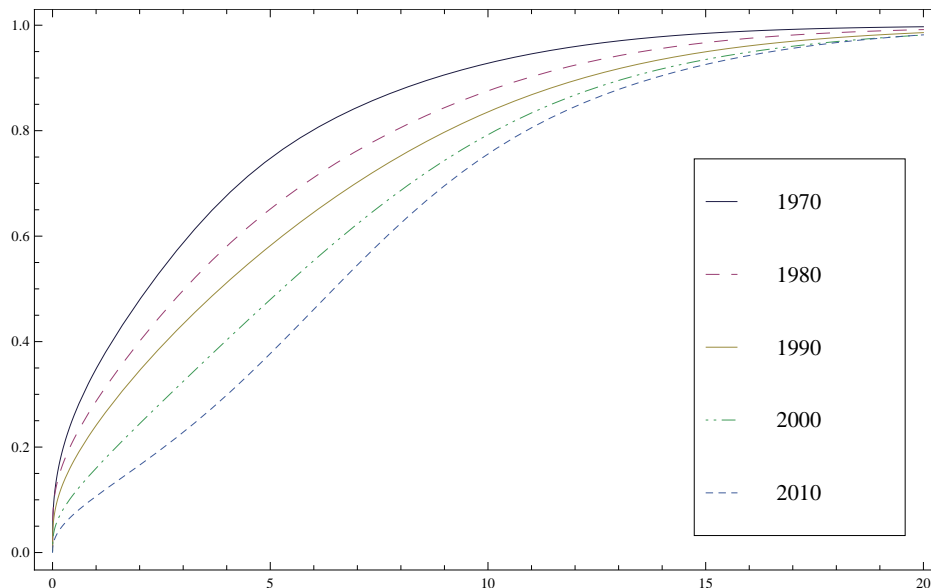
Figure 2: Global CDF of years of schooling

concur that inequality levels decreased independently of their sensitivity to the upper tail.

To quantify the reduction in inequality, Table 4 presents GE measures that (i) are more sensitive to less educated individuals (the MLD - $\theta = 0$), (ii) give more importance to higher education ($\theta = 2$) or (iii) weight equally all parts of the distribution (the Theil index - $\theta = 1$). In line with previous studies, our estimates reveal a substantial decrease in global inequality since the 1970s. This outstanding reduction can be attributable to the decrease of the illiterate population (Morrisson and Murtin, 2013), although there could be other relevant factors affecting this reduction in educational disparities, such as convergence in compulsory years of schooling (Murtin and Viarengo, 2011) and the expansion of primary education in developing countries, whose attainment rates reached 90 percent in 2012 (United Nations, 2014). Notwithstanding the overall reduction, if we attach more weight to illiterate population, inequality levels tend to increase over the 1970s. The Theil index and the GE(2) measure, in contrast, present a decreasing trend during the whole period, which is consistent with those results reported in Benaabdelaali et al. (2012). As pointed out by our estimates, inequality measures might exhibit opposite trends depending on the weighting structure. Then, when there is no Lorenz dominance, the consideration of inequality measures that weights differently each part of the distribution is of main interest.

We exploit the property of decomposability by population subgroups of the GE measures
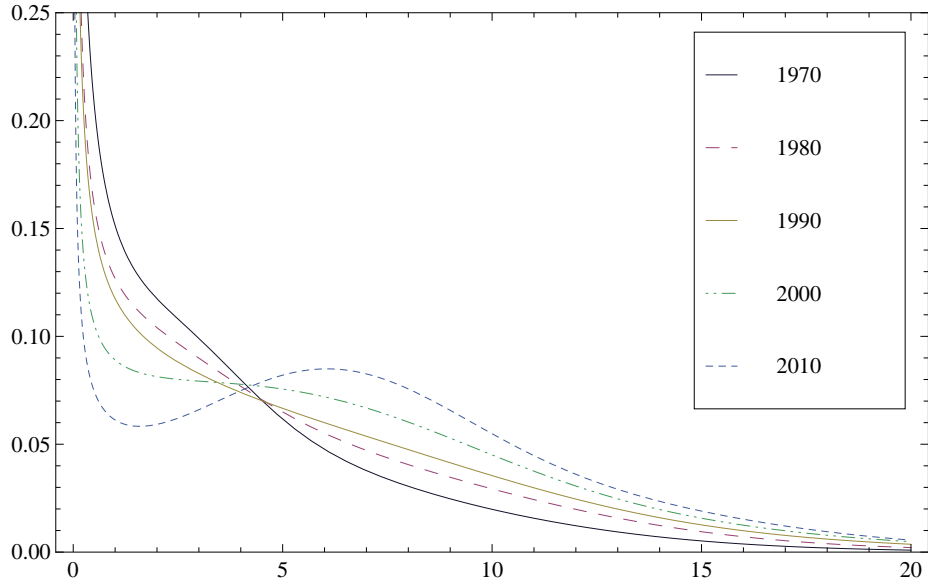
18

Figure 3: Global PDF of years of schooling

Table 4: Global inequality in education (1970-2010)

Year	1970	1975	1980	1985	1990	1995	2000	2005	2010
MLD	1.8084	1.8806	1.6931	1.3579	1.1535	0.9921	0.8485	0.6736	0.5691
Between	0.1978	0.1713	0.1414	0.1180	0.1047	0.0910	0.0806	0.0709	0.0624
Within	1.6106	1.7093	1.5517	1.2399	1.0488	0.9011	0.7678	0.6027	0.5067
Theil	0.5646	0.5441	0.5030	0.4740	0.4509	0.4000	0.3577	0.3068	0.2756
Between	0.1972	0.1729	0.1418	0.1195	0.1062	0.0903	0.0796	0.0693	0.0614
Within	0.3674	0.3712	0.3612	0.3545	0.3447	0.3098	0.2781	0.2375	0.2142
GE (2)	0.6227	0.5787	0.5140	0.4804	0.4571	0.3919	0.3457	0.2910	0.2633
Between	0.2288	0.1995	0.1597	0.1334	0.1177	0.0972	0.0847	0.0726	0.0643
Within	0.3939	0.3792	0.3543	0.3470	0.3394	0.2948	0.2611	0.2184	0.1989

to break down overall inequality into differences in MYS across countries and disparities in educational outcomes within countries. While differences between countries decreased rapidly over the whole period, the fall of within-country inequality is less pronounced. Our estimates also reveal that the differences in education within countries played a predominant role in global inequality, while differences across countries represented a reduced proportion. Moreover, the share of this component in overall inequality steadily decreased over the last 40 years, representing less than 25 percent in 2010. This finding suggests that previous studies on the evolution of international inequality in mean levels of education only analyzed a residual proportion of the global disparities.

In summary, these major advances in educational attainment over time may well reflect
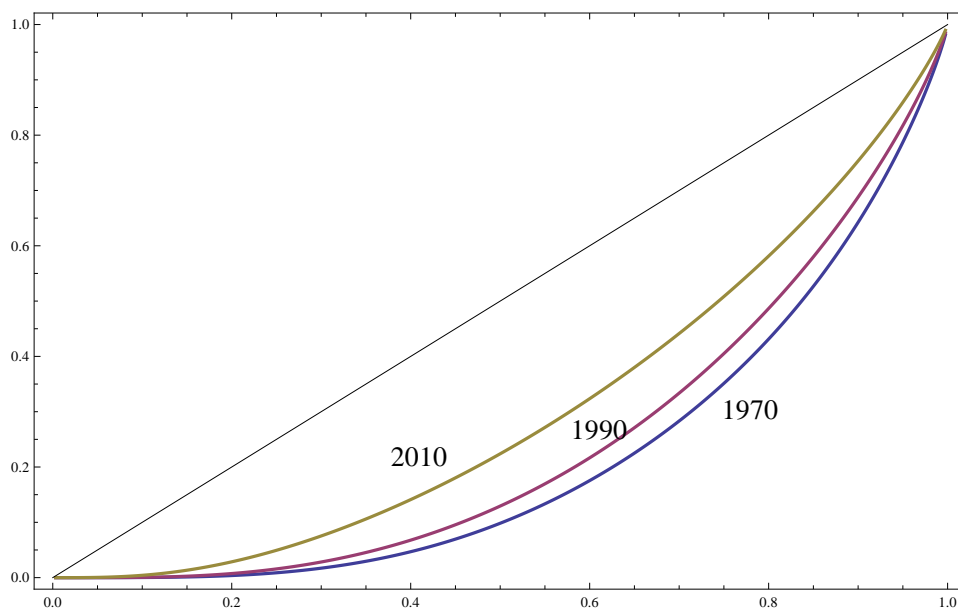
Figure 4: Global Lorenz curves of education

those efforts made by countries and international organizations to increase workforce skills and to make progress towards the achievement of universal primary education, particularly since the World Declaration on Education for All, adopted by UNESCO in 1990, and the formulation of the Millennium Development Goals (Masino and Niño-Zarazúa, 2016). It should be acknowledged, however, that educational attainment and inequality measures may be adversely affected by data quality and comparability issues. These potential (and critical) problems have been flagged by several educational researchers (see, for example, Behrman and Rosenzwei, 1994; Bloom, 2007; De la Fuente and Domenech, 2006; Krueger and Lindahl, 2001).

In particular, educational data may be problematical not only because they may be affected by those well-known problems afflicting data such as censuses and household surveys, but also because of issues related to comparability across countries and over time within a country (Behrman and Rosenzwei, 1994: 148). For example, the quality (and quantity) of international data on education could have increased over time due to, among other reasons, capacity building and training efforts to improve statistical skills, particularly in developing countries (Lievesley, 2001). It is conceivable, therefore, that these potential improvements in data gathering and quality may have contributed to the large progress in educational attainment reported here. A second, major, concern refers to the purely quantitative nature of our educational attainment measure, i.e. the average years of

schooling, which does not take into account potential differences in the quality of education across countries. Ignoring quality differences in education may lead, from a comparative perspective, to a distorted picture of how countries progress towards better educational outcomes.

To conclude this overview of our new educational attainment and inequality estimates, it is hence important to recognize that the data issues discussed above suggest that one should be cautious in drawing very narrow conclusions derived from MYS estimates. Nonetheless, the substantial increase in years of schooling reported in this section, along with the substantial decrease in educational inequality, suggest that, even considering the existence of potential noise affecting MYS estimates, countries are, on average, progressing towards better educational outcomes.

4   Comparison with previous estimates

In this section, we compare our estimates of MYS and educational inequality measures with those series provided by BL and Benaabdelaali et al. (2012) respectively. Starting with the comparison with BL, Table 5 presents the average and the standard deviation of MYS estimates using the continuous approach, along with BL data (the discrete approach) over the period 1970-2010. Although the correlation between both set of estimates is fairly high, both in levels and in first differences, MYS figures from BL data are considerably higher in all cases.[10]

It should be acknowledged that using overestimated levels of MYS can affect empirical estimates of the role of education on a variety of socio-economic aspects, particularly if this potential overestimation is not of equal size across all countries. The gap between the discrete and the continuous approach differs across countries, so it is not possible to figure out, *a priori*, how the estimates of the effect of education could be affected.

We now check these potential differences through the estimation of a simple production function. It should be highlighted from the outset that our aim is not to analyse those

---

[10]Following Krueger and Lindahl (2001), and Cohen and Soto (2007), this correlation between both series of estimates may be an indication about the reliability of our estimates. In our case, even 5- and 10-year differences are strongly correlated, which would reinforce the validity of our estimates. It could be argued, however, that the strong level of correlation between both series is driven by the use of the same data on educational attainment, which can correlate positively the measurement errors of both series. While it might foster correlation levels, the use of the same data on educational attainment may not explain itself the high correlation in first differences (Cohen and Soto, 2007) because the method of estimation differs, and even the conceptualization of education is different.

Table 5: Comparison of BL data on average years of schooling (discrete) with the theoretical values of MYS for the GG distribution (continuous). Population over-15.

| | | Levels | | Correlation | |
		Discrete	Continuous	levels	10-year D.
Total	1970	4.3618 (2.6103)	3.3018 (1.9772)	0.9138	-
	1980	5.4697 (2.7185)	4.3103 (2.1893)	0.9316	0.8888
	1990	6.4832 (2.6976)	5.2677 (2.3026)	0.9449	0.8153
	2000	7.4143 (2.7767)	6.2596 (2.5757)	0.9562	0.7813
	2010	8.3970 (2.7792)	7.3504 (2.8311)	0.9603	0.7372
Males	1970	4.8816 (2.5729)	3.7796 (2.0358)	0.9115	-
	1980	5.9923 (2.6170)	4.8246 (2.2055)	0.9267	0.8349
	1990	6.9564 (2.5407)	5.7356 (2.2462)	0.9378	0.7925
	2000	7.7970 (2.5951)	6.6253 (2.4768)	0.9485	0.7839
	2010	8.6425 (2.5983)	7.5870 (2.7007)	0.9520	0.7174
Females	1970	3.8670 (2.7005)	2.8657 (2.0102)	0.9282	-
	1980	4.9549 (2.8669)	3.8297 (2.2647)	0.9368	0.9183
	1990	6.0470 (2.9126)	4.8485 (2.4537)	0.9496	0.8655
	2000	7.0665 (3.0177)	5.9554 (2.7808)	0.9624	0.7509
	2010	8.1739 (3.0078)	7.1803 (3.0489)	0.9652	0.7255

Notes: Standard deviations in parenthesis.

factors affecting income levels, but just to assess whether regression coefficients for MYS differ across both set of estimates (i.e. the discrete and continuous approaches). To do so, we follow a similar approach to Cohen and Soto (2007) and estimate the following aggregate production function for country i at time t:

$$Y_{it} = A_{it}K_{it}H_{it} \qquad (10)$$

where Y_{it} is aggregate income or gross domestic product (GDP), A_{it} is the total factor productivity (TFP), K_{it} is physical capital and $H_{it} = h_{it}L_{it}$ (h_{it} denotes human capital per worker; L_{it} represents labour force). Taking logs of (10), diving by L_{it}, and assuming that both $log(h_{it})$ and A_{it} are represented, respectively, by the Mincerian approach to human capital[11], and by the sum of a fixed effect, time effect and the remaining error term, we derive the following equation for the log of GDP per worker in country i at time t:

$$log(y_{it}) = \alpha log(k_{it}) + \beta mys_{it} + \mu_i + \lambda_t + \varepsilon_{it} \qquad (11)$$

where mys_{it} is the itth observation of MYS; μ_i denotes country specific effects; λ_t represents the specific time effect (common to all countries) and ε_{it} the remainder disturbance term. In order to compare both approaches to estimate MYS, we constructed an unbalanced

---

[11]Although the macro Mincer (1974) approach assumes that the logarithm of human capital depends linearly on years of schooling and cumulative work experience, we follow here Krueger and Lindahl (2001) and Cohen and Soto (2007), and have suppressed the experience term.

panel of 132 countries from 1990 to 2010 (five year periods). Information regarding income *per* worker and physical capital *per* worker was obtained from the Penn World Tables, version 8.1 (see Feenstra et al., 2015, for a description of the data). We first estimate Eq(11) by means of a simple fixed-effects (FE) model. It is conceivable, however, that fixed-effects estimations may result in biased estimates due to the potential endogeneity of physical and human capital. Therefore, to address these potential endogeneity issues, we propose to complement our estimations with a two-stage least squares (2SLS) approach and the generalized method of moments (GMM), using the first two lags of the potential endogenous variables as instruments.

Table 6 reports FE, 2SLS and GMM estimates for the proposed production function. Starting with the FE specification, the parameter estimate for MYS using our estimates is considerably larger than the parameter for MYS using BL data set. These results confirm our expectations about potential differences across estimations using the discrete and continuous approaches. In addition, the standard error for MYS using our series is lower than the corresponding model using BL data. We interpret this as a baseline indication that our series may provide more accurate estimates of the effect of MYS on per capita income levels. However, we argued before that the relationship between income levels, physical capital and human capital may be affected for endogeneity issues. Hence, to address this potential source of bias we also report results using 2SLS and GMM estimators. Although the coefficients associated to MYS are slightly smaller when attempting to correct for endogeneity, our previous findings hold with both instrumental variable approaches. [12]

Moving now to education inequality measures, Table 7 presents a comparison of our estimates with those provided by Benaabdelaali et al. (2012), which is the most recent data set on educational inequality, covering the period 1950-2010. Their estimates are also computed using BL data on attainment ratios and consider changes in the duration system of the countries. To the best of our knowledge, existing series on educational inequality only provide Gini coefficient estimates, hence the following comparison focuses solely on this inequality measure.[13]

To analyze the shape of the distribution of the educational Gini indices across countries,

---

[12] Kleibergen-Paap (KP) under-identification and KP weak identification tests suggest that the selected instruments are relevant, and Hansen over-identification tests suggest that those instruments are valid (see Table 6).

[13] Morrisson and Murtin (2013) also computed Theil indices to analyze global inequality in education. They used, however, data for 32 macro-regions, which prevent us from comparing both sets of estimates on this occasion.

Table 6: Comparison of aggregate production function panel estimates between the discrete and continuous approaches

| | FE | | 2SLS | | GMM | |
	Discrete	Continuous	Discrete	Continuous	Discrete	Continuous
log (k)	0.564	0.552	0.457	0.448	0.468	0.450
	(0.043)	(0.042)	(0.039)	(0.039)	(0.039)	(0.039)
mys	0.041	0.077	0.038	0.071	0.035	0.067
	(0.020)	(0.015)	(0.019)	(0.014)	(0.019)	(0.014)
Observations	1070	1070	808	808	808	808
R-squared	0.6317	0.6512	0.6262	0.6387	0.627	0.6391
K-P LM			126.21	111.174	126.21	111.174
K-P Wald F			111.425	138.546	111.425	138.546
Hansen J			4.825	4.506	4.825	4.506

Notes: Dependent and independent variables taken at the beginning of the five-year period. Following Cohen and Soto (2007) mys refers to MYS of population aged 25 years and older. All models include time dummies. Robust standard errors reported in parentheses. Panel 2SLS and GMM models computed with the Stata command *xtivreg2* developed by Schaffer (2012).

we present the simple average of the Gini index along with the standard deviation in parenthesis. As expected, the Gini index is higher using the continuous approach, since the discrete approach ignores differences in educational outcomes within levels. Interestingly, the gap between both estimates increases over time; during the 1970s, the 'continuous' Gini was only two points higher than the Gini index obtained using the discrete approach, while this gap rose to seven points during the last decade.

Table 7: Comparison of theoretical and empirical Gini indices of education

| | Year | Levels | | Correlation | |
		Discrete	Continuous	levels	10-year D.
	1970	0.5283 (0.2286)	0.5537 (0.1979)	0.9604	-
	1980	0.4700 (0.2114)	0.5135 (0.1784)	0.9586	0.6550
over-15	1990	0.4132 (0.1887)	0.4736 (0.1591)	0.9528	0.8030
	2000	0.3568 (0.1779)	0.4250 (0.1603)	0.9376	0.5742
	2010	0.3089 (0.1585)	0.3755 (0.1525)	0.8914	0.4070
	1970	0.5698 (0.2420)	0.5861 (0.2117)	0.9492	-
	1980	0.5174 (0.2285)	0.5616 (0.1888)	0.9526	0.6218
over-25	1990	0.4581 (0.2079)	0.5215 (0.1679)	0.9560	0.7692
	2000	0.3923 (0.1966)	0.4647 (0.1705)	0.9425	0.6449
	2010	0.3382 (0.1811)	0.4132 (0.1677)	0.8984	0.3291

The correlation coefficient between both sets of estimates is fairly high in levels, being above 90 percent during the whole period, but this correlation is weaker in 10-year differences. We must be cautious about the high correlation observed between both measures, because educational Gini indices provided by Benaabdelaali et al. (2012) are computed

using BL data on attainment rates. As a consequence, correlation coefficients may be overestimated because of the positive correlation between the errors of both measures.
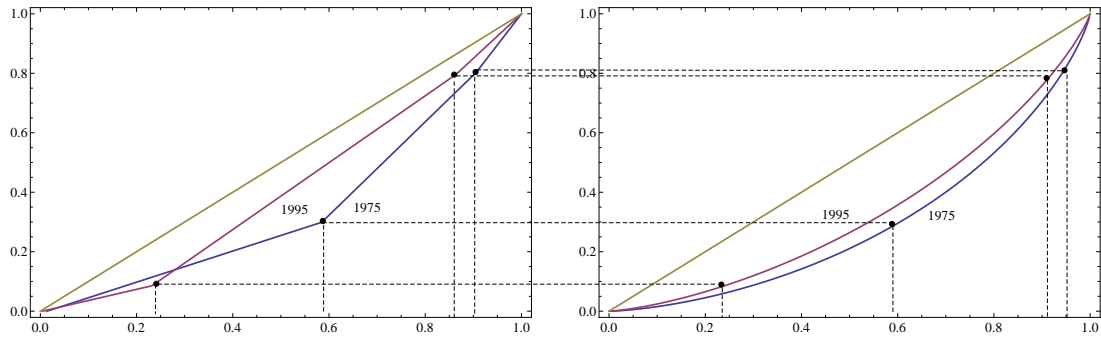

Figure 5: Lorenz curves of education in Latvia using the discrete and the continuous approaches

The comparisons of MYS and inequality measures between the continuous methodology and the discrete approach suggest that the results differ depending on the data used. To further illustrate this point, Figure 5 shows the educational Lorenz curves for Latvia in 1975 and 1995. The Lorenz curve of the discrete approach is presented in the left panel. These curves are constructed by plotting on the horizontal axis the accumulated attainment rates and, on the vertical axis, the cumulative proportion of education, which is given by the weighted sum of the attainment rate and the official duration of each level, divided by overall MYS. Following Castello and Domenech (2008), we do not break down educational levels into complete and incomplete, thus assuming that all individuals classified in some educational level have completed it and hence we assign them the official duration. By applying this method we obtain a step-curve truncated at the horizontal axis, due to the assumption that illiterate people have zero years of schooling. On the right side, we plot the Lorenz curves using the continuous methodology developed in this paper. It should be noted that, by construction, both methodologies produce curves that pass through the black points, as highlighted in the graph by the dashed lines. However, Lorenz dominance is only observed using the continuous approach. These two opposite conclusions are the consequence of assuming different patterns as regards the incomplete levels of education, which lead to different curvatures between the aforementioned black points.

5   Conclusions

In this paper, we present new estimates for MYS and inequality measures of education, covering 142 countries over the 1970-2010 period, with the aim of providing more accurate estimates than existing series. Previous attempts to measure educational attainment computed MYS as a weighted average of the official duration of each level and attainment rates. Educational cycles were divided into complete and incomplete, whether the educational level had been finished or not. The main shortcoming of these previous studies is the assumption that all individuals who have not completed a particular educational level have the same arbitrary number of years of schooling. This implicit choice may bias the estimates of MYS and, in addition, the direction of the bias cannot be determined *a priori*. Even when the distribution within each educational stage is unknown, the accuracy of the estimates on MYS and inequality measures rely on a good approximation of the unidentified segments of the empirical distribution.

In this study, we deploy a flexible parametric model to estimate the distribution of years of schooling. This methodology allows us to impose more reliable assumptions about the differences in years of schooling within each level of education. In addition, to assess inequality levels in educational outcomes, we focus on the GE family of inequality measures. Different parameter values have been considered, thus changing the sensitivity of these indicators to the upper tail. The use of additively decomposable measures is more convenient than relying solely on Gini coefficients for practical purposes. Indeed, the main advantage of our estimates is that overall inequality can be computed for any group of countries using only the information included in our data set.

After briefly presenting and discussing main trends on educational attainment and educational inequality over the past decades, we compare our estimates with those based on a discrete approach to measure educational attainment and education inequality. First, we find that average educational outcomes estimated using the discrete approach, such as Barro and Lee (2013), are substantially higher than our estimates. In addition, although the correlation between both sets of estimates is fairly high, we observe that our series may provide more accurate estimates than those relying on the discrete approach. Similar conclusions can be drawn when comparing our educational Gini coefficients and those provided by Benaabdelaali et al. (2012), although the correlation between these series in first differences is substantially lower.

In sum, we show that the results of the discrete approach seem to be extremely sensitive to the assumptions made on the number of years of schooling assigned to the incomplete levels. In contrast, the methodology we deploy here avoids relying on such assumptions about the unknown parts of the distribution, which are estimated using a flexible parametric model. These improved series on educational attainment and education inequality may be useful to improve our understanding of the role of education on different aspects of development.

References

Acemoglu, D. and Angrist, J. (2001). How large are human-capital externalities? evidence from compulsory-schooling laws. In *NBER Macroeconomics Annual 2000, Volume 15*, pages 9–74. MIT PRess.

Arulampalam, W., Naylor, R. A., and Smith, J. P. (2004). A hazard model of the probability of medical school drop-out in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(1):157–178.

Baker, D. P. and Holsinger, D. B. (1996). Human capital formation and school expansion in asia: Does a unique regional model exist. *International Journal of Comparative Sociology*, 37(1-2):159–173.

Barro, R. J. and Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics*, 104:184–198.

Becker, G. S., Murphy, K. M., and Tamura, R. (1990). Human capital, fertility, and economic growth. *Journal of Political Economy*, 98(5):12–37.

Behrman, J. R. and Rosenzweig, M. R. (1994). Caveat emptor: Cross-country data on education and the labor force. *Journal of Development Economics*, 44(1):147 – 171.

Benaabdelaali, W., Hanchane, S., and Kamal, A. (2012). Educational inequality in the world, 1950–2010: Estimates from a new dataset. *Research on Economic Inequality*, 20:337–366.

Bloom, D. E. (2007). Measuring global educational progress. In Cohen, J. E., Bloom,

D. E., and Malin, M. B., editors, *Education for All: A Global Agenda*, pages 33–120. American Academy of Arts and Sciences, Cambridge: MA.

Castell-Climent, A. (2010). Channels through which human capital inequality influences economic growth. *Journal of Human Capital*, 4(4):394–450.

Castelló-Climent, A. (2008). On the distribution of education and democracy. *Journal of Development Economics*, 87(2):179–190.

Castelló-Climent, A. and Doménech, R. (2008). Human capital inequality, life expectancy and economic growth*. *The Economic Journal*, 118(528):653–677.

Chotikapanich, D., Griffiths, W. E., Prasada Rao, D., and Valencia, V. (2012). Global income distributions and inequality, 1993 and 2000: Incorporating country-level inequality modeled with beta distributions. *Review of Economics and Statistics*, 94(1):52–73.

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

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

De la Fuente, A. and Domenech, R. (2013). Cross-country data on the quantity of schooling: a selective survey and some quality measures. *BBVA Working Papers*, (13/27).

Feenstra, R., Inklaar, R., and Timmer, M. (2015). The next generation of the penn world table. *American Economic Review*, 105(10):3150–3182.

Filmer, D. and Pritchett, L. (1999). The effect of household wealth on educational attainment: evidence from 35 countries. *Population and Development Review*, 25(1):85–120.

Gomes, O., Combes, C., and Dussauchoy, A. (2008). Parameter estimation of the generalized gamma distribution. *Mathematics and Computers in Simulation*, 79(4):955–963.

Gregorio, J. D. and Lee, J.-W. (2002). Education and income inequality: new evidence from cross-country data. *Review of Income and Wealth*, 48(3):395–416.

Hannum, E. and Buchmann, C. (2005). Global educational expansion and socio-economic development: An assessment of findings from the social sciences. *World Development*, 33(3):333 – 354.

Jenkins, S. P. (2009). Distributionally-sensitive inequality indices and the gb2 income distribution. *Review of Income and Wealth*, 55(2):392–398.

Jordá, V., Sarabia, J. M., and Prieto, F. (2014). On the estimation of the global income distribution using a parsimonious approach. *Research on Economic Inequality*, 22:115–145.

Krueger, A. B. and Lindahl, M. (2001). Education for growth: Why and for whom? *Journal of Economic Literature*, 39(4):1101–1136.

Lewin, K. M. (2009). Access to education in sub-saharan africa: Patterns, problems and possibilities. *Comparative Education*, 45(2):151–174.

Lievesley, D. (2001). Making a difference: A role for the responsible international statistician? *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(4):367 406.

Lin, J. Y. (2015). The Washington Consensus revisited: a new structural economics perspective. *Journal of Economic Policy Reform*, 18(2):96–113.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3 – 42.

Lutz, W., Goujon, A., KC, S., and Sanderson, W. (2007). Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000. *Vienna yearbook of population research*, 5:193–235.

Marshall, A. W. and Olkin, I. (2007). *Life Distributions. Structure of Nonparametric, Semiparametric, and Parametric Families*. New York: Springer.

Masino, S. and Niño-Zarazúa, M. (2016). What works to improve the quality of student learning in developing countries? *International Journal of Educational Development*, 48:53 – 65.

Meschi, E. and Scervini, F. (2013). A new dataset on educational inequality. *Empirical Economics*, 47(2):695–716.

Mincer, J. (1974). *Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2.* Columbia University Press, New York.

Morrisson, C. and Murtin, F. (2009). The century of education. *Journal of Human Capital*, 3(1):1–42.

Morrisson, C. and Murtin, F. (2013). The kuznets curve of human capital inequality: 18702010. *The Journal of Economic Inequality*, 11(3):283–301.

Murtin, F. and Viarengo, M. (2011). The expansion and convergence of compulsory schooling in western europe, 1950–2000. *Economica*, 78(311):501–522.

Oreopoulos, P. and Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *The Journal of Economic Perspectives*, 5(1):159–184.

Plank, S., DeLuca, S., and Estacion, A. (2005). Dropping out of high school and the place of career and technical education: A survival analysis of surviving high school. *National Research Center for Career and Technical Education.*

Psacharopoulos, G. (1994). Returns to investment in education: A global update. *World Development*, 22(9):1325 – 1343.

Psacharopoulos, G. and Arriagada, A. M. (1986). Educational composition of the labour force: An international comparison. *International Labour Review*, 12(5):561–574.

Ram, R. (1990). Educational expansion and schooling inequality: International evidence and some implications. *The Review of Economics and Statistics*, 72(2):266–274.

Rebelo, S. (1991). Long-run policy analysis and long-run growth. *Journal of Political Economy*, 99(3):500–521.

Sarabia, J. M. and Jordá, V. (2014). Explicit expressions of the pietra index for the generalized function for the size distribution of income. *Physica A: Statistical Mechanics and its Applications*, 416:582–595.

Sarabia, J. M., Jordá, V., and Remuzgo, L. (2016). Computation of the theil indices in parametric families of income distributions. *Review of Income and Wealth*, forthcoming.

Schaffer, M. E. (2012). xtivreg2: Stata module to perform extended iv/2sls, gmm and ac/hac, liml and k-class regression for panel data models. *Statistical Software Components*.

Sen, A. (1988). The concept of development. *Handbook of development economics*, 1:9–26.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.

Stiglitz, J., Sen, A., and Fitoussi, J. (2010). *Mismeasuring Our Lives: Why GDP Doesn't Add Up*. New Press, New York.

Thomas, V., Wang, Y., and Fan, X. (2001). *Measuring education inequality: Gini coefficients of education*, volume 2525. World Bank Publications.

UNESCO (2013). UIS methodology for estimation of mean years of schooling. Available at: http://www.uis.unesco.org/Education/Documents/mean-years-schooling-indicator-methodology-en.pdf.

UNESCO (2014). Statistical yearbook. *Paris*.

United Nations (2014). *The Millennium Development Goals Report 2008*. United Nations Publications.

World Bank (2005). *World Development Report 2006: equity and development*. World Bank.