J Econ Growth (2007) 12:51–76
DOI 10.1007/s10887-007-9011-5

Growth and human capital: good data, good results

Daniel Cohen · Marcelo Soto

Published online: 27 March 2007

Abstract We present a new data set for years of schooling across countries for the 1960–2000 period. The series are constructed from the OECD database on educational attainment and from surveys published by UNESCO. Two features that improve the quality of our data with respect to other series, particularly for series in first-differences, are the use of surveys based on uniform classification systems of education over time, and an intensified use of information by age groups. As a result of the improvement in quality, these new series can be used as a direct substitute for Barro and Lee's (2001; Oxford Economic Papers, 3, 541–563) data in empirical research. In standard cross-country growth regressions we find that our series yield significant coefficients for schooling. In panel data estimates our series are also significant even when the regressions account for the accumulation of physical capital. Moreover, the estimated macro return is consistent with those reported in labour studies. These results differ from the typical findings of the earlier literature and are a consequence of the reduction in measurement error in the series.

Keywords Human capital · Education · Economic growth

JEL Classification O10

1 Introduction

The role of human capital in economic growth is an ongoing topic that has changed course at least three times over the past two decades. The idea that human capital

Daniel Cohen
Paris School of Economics, OECD Development Centre and CEPR. 48 Boulevard Jourdan, 75014 Paris, France

Marcelo Soto (✉)
Instituto de Análisis Económico Campus UAB, 08193 Bellaterra, Barcelona, Spain

could generate long-term sustained growth was one of the critical features of the "new growth" literature initiated by Lucas (1988) and Romer (1990). Later on, a neo-classical revival evolved from Mankiw Romer and Weil (1992), who described human capital as an ordinary input unable to generate endogenous growth. Building upon this neo-classical approach, a new "revisionist" view gained influence, following the studies of Benhabib and Spiegel (1994), Pritchett (2001), and Bils and Klenow (2000), according to which the role of human capital in economic growth has been vastly overstated.

We argue in this paper that part of the reason why the debate erred between these extremes is due to the measurement of human capital, both conceptually and empirically. Conceptually, there has not been a clear-cut definition on how human capital should be represented. Years of schooling have long been considered as a good proxy. Yet a simple glance at the data shows that the regions where the growth rate of schooling has been the fastest are also those where it started from very low levels (Africa being a prime example). It is hard to believe that a country that increased its average years of schooling from 1 to 2 really doubled its stock of human capital and should therefore, eventually, double its output. In the case of Mankiw Romer and Weil (1992), human capital is indirectly represented through a law of motion similar to the process followed by physical capital. In their model, a fraction of GDP (which they assume proportional to secondary school enrolment) is diverted towards raising human capital. Yet, as demonstrated in Cohen (1996), this formulation is clearly rejected by the data. It is only recently that the macro-literature has turned to the micro-literature, specifically in terms of the Mincerian approach to human capital, to redefine the link between schooling and human capital. According to this approach human capital is an exponential function of the years of schooling, which results in a log-linear (instead of a log-log) correspondence between income and years of schooling. According to the Mincerian representation, the poor countries' increase in human capital is modest: as we show below the gap in human capital between rich and poor countries has been constant over the last forty years.

The second problem faced by empirical studies is related to the quality of the data itself. This critical problem has recently been emphasized by De la Fuente and Domenech (2002, 2006). Focusing on a subgroup of 21 OECD countries, they have demonstrated that the existing human capital data is fairly unreliable. Measurement errors are also emphasized by Krueger and Lindahl (2001), who show that there is little information in the data on years of schooling used in the growth regressions reported by Benhabib and Spiegel (1994) and Pritchett (2001).

This paper's contribution resides in a new effort to raise the quality of the data on human capital for a broad group of countries. These new series are based on the OECD database on educational attainment described below. To calculate the years of schooling in countries not covered by the OECD database we have used censuses and historical information published by UNESCO and other sources. One critical characteristic in our methodology, which we explain in detail in Sect. 2, is that we exploit the information available by age groups. This allows us to rely more on observed data and less on assumptions. This has not been done before. Another important feature of our methodology is that we avoid the simultaneous use of censuses in a particular country that are based on different classifications of levels of education. Not accounting for variations in the classification systems would introduce noise to the measured changes in schooling between two different dates. We argue and show in Sect. 3 that these two features lead to a significant reduction in the measurement error in the

data. In Sect. 4 we analyze the performance of these new series in standard growth regressions. Unlike the earlier empirical literature, we obtain coefficients that are both positively and significantly associated to our schooling variable. Finally, in Sect. 5, we turn to more recent techniques for panel data in order to estimate a simple production function, modified to incorporate the Mincerian approach to human capital. We show that the estimated coefficients for schooling closely match the average return found in micro studies.

2 A new data set

2.1 Methodology

As in most of the earlier literature on human capital measurements, we build the average number of years of schooling in a country by multiplying the population's shares of educational attainment by the appropriate length (in years) of each educational category (i.e. primary, secondary and higher education). The length may vary from country to country, which is taken into account in this paper.

The difference between our approach and earlier measurements resides in the fact that we make use of the information on educational attainment by age. This information has not been exploited before. To achieve this, three main sources are used: (i) the OECD database on education; (ii) national censuses or surveys published by UNESCO's Statistical Yearbook and the Statistics of educational attainment and illiteracy; and (iii) censuses obtained directly from national statistical agencies' web pages. More details about these sources can be found at http://www.iae-csic.uab.es/soto/data.htm.

A second feature of our methodology is that it seeks to keep the series consistent for a particular country over time by avoiding the use of censuses based on different classification systems of education. If alterations to the classification systems are not accounted for properly, the changes in measured years of schooling between two dates would be affected by these alterations. This is a major cause of measurement error. Whenever we identify a change in the durations,[1] we exclusively use the census information based on the latest classification and apply the backward extrapolation described below to calculate the years of schooling for earlier dates.

Based on reports from its member and fifteen non-member countries, the OECD has collected detailed information on educational attainment, starting from the end of the 1980s. This information refers to the population aged 15 and above and is broken down into different age groups. Summary statistics on the labour force's school attainment in these countries can be found in the OECD's periodicals "Education at a glance: OECD indicators" and "Investing in Education: Analysis of the world education indicators". The main advantage of the OECD database is that it is presented in a standardized form across countries and over time. In this paper we attempt to extend the OECD database to additional periods and countries.

The methodology employed is the following. We first calculate the educational attainment of the population by five-year age groups for each of the years 1960, 1970, 1980, 1990 and 2000 using OECD and other censuses on schooling. For any year t for

[1] We use UNESCO's Statistical Yearbooks to identify these alterations.

which there is a census we calculate the number of years of schooling of population aged 15 and above (ys_t) as a weighted average of different age groups:[2]

$$ys_t = \sum_{g=1}^{G} l_t^g ys_t^g \tag{1}$$

where l_t^g represents the population share of group g in population 15 and above, obtained from the United Nations Demographic Yearbook; ys_t^g is the number of years of schooling of group g; $g = 1$ is the 15–19 age group; $g = 2$, the 20–24 age group, and so on until the oldest age group available $g = G$ (typically 65 and above). ys_t^g is given by (we omit the time subscript),

$$ys^g = \sum_j a_j^g D_j$$

where a_j^g is the fraction of group g having attained educational level j and D_j is the corresponding duration in years.

When no census is available for a date before t we make a backward extrapolation under the assumption that $ys_{t-5}^g = ys_t^{g+1}$ for $g = 3$ to $G - 1$. That is, we assume that after a cohort reaches the age 25 its average number of years of schooling remains unchanged. This assumption makes it possible to infer the years of schooling of the cohorts aged 25 and above at date $t - 5$ from the census of date t. Note that no inference can be made for the two youngest and the oldest groups in $t - 5$ because their schooling level is not directly observed at date t. The years of schooling of these three age groups are calculated from enrolment data as explained below. Then, with the estimates of years of schooling for all the G age groups at hand, we calculate ys_{t-5} as a weighted average across age groups as follows:

$$ys_{t-5} = \sum_{g=1}^{2} l_{t-5}^g ys_{t-5}^g + \sum_{g=3}^{G-1} l_{t-5}^g ys_t^{g+1} + l_{t-5}^G ys_{t-5}^G \tag{2}$$

The calculations for earlier dates are carried out by recursively applying the same procedure. This methodology makes it possible to eventually calculate the years of schooling in 1960 from a 1990 census.

Next, if no census is available for a date after t, we apply the same methodology to make a forward extrapolation. More precisely, we assume that $ys_{t+5}^g = ys_t^{g-1}$ for $g = 4$ to G. Note that in the forward procedure none of the first three age groups' schooling levels can be inferred from the census of year t since part of the population aged 20–24 and below in year t may have had some schooling between dates t and $t+5$ (as in the backward extrapolation, we ignore the studies made at age 25 and above). Consequently the years of schooling of these groups need to be calculated separately from enrolment data as explained shortly. Then we calculate ys_{t+5} according to:

$$ys_{t+5} = \sum_{g=1}^{3} l_{t+5}^g ys_{t+5}^g + \sum_{g=4}^{G} l_{t+5}^g ys_t^{g-1} \tag{3}$$

Estimations for later dates are carried out by recursively applying (3).

The years of schooling of cohorts whose school attainment is not directly observed from a census are calculated from enrolment data. The following example illustrates how we proceed. Suppose that we need to calculate the educational attainment for

[2] Here, we describe the methodology to construct the years of schooling of population 15 and above. We use the same approach to estimate the schooling of other age groups used in the literature, namely the population aged 25 and above or the population aged 15–64.

the 60–64 age group in 1980. Assuming that the age of entry to primary education is six, this cohort was of the age to start primary education between the years 1922 and 1926. By calculating the ratio of new entrants into primary school first grade to the 6-year-old population—i.e. the net intake rate—during, for instance, 1924, it is possible to obtain an estimate of the fraction of that cohort that attended primary school. The same procedure provides an estimate of attendance of upper school categories (i.e. secondary and higher education) for cohorts with no information available from censuses. Since the number of students enrolled in each category cannot be used directly to calculate the number of new entrants (because of repeaters, dropouts and pupil growth), they need to be adjusted.[3] Appendix A.1 describes how we calculate the net intake rates from enrolment data. Several sources are used to collect enrolment data. The main source is Mitchell (1993, 1998a,b), who has published historical series on school enrolment by educational category for most countries, starting from the second half of the 19th century. This information is combined with UNESCO's Statistical Yearbook, which starting from 1950 also systematically publishes data on enrolment by category. In general both sources coincide, but this is not always the case. Therefore for years after 1950, UNESCO's data is used. Population tables by age are from Mitchell, the United Nations Demographic Yearbook, the U. S. Census Bureau and national agencies.

Nehru, Swanson and Dubey (NSD, 1995) have already used Mitchell's series to build educational indexes. However their series have not enjoyed success in empirical work because they make no use of information from censuses. Indeed, some of their series bear little relationship with data measured directly from censuses. Moreover, De la Fuente and Domenech (2006) have noted the incidence of some implausible results in NSD's database.[4] One important difference between NSD's approach and the present approach is that here enrolment data is only used to fill in missing information rather than to construct the entire database.

There are at least two major caveats in our methodology. First, it is assumed that the mortality rate is distributed homogeneously within each age group, while it could be argued that the more educated people have lower mortality rates. Under these circumstances, the actual years of schooling would be such that $ys^g_{t-5} < ys^{g+1}_t$ and $ys^g_{t+5} > ys^{g-1}_t$, which means that the backward and forward procedure over and underestimate the years of schooling, respectively. The lack of information on mortality rates by educational attainment precludes a more accurate method of estimation.[5] Yet, it is important to stress that our approach does account for differences

[3] If repeaters, dropouts and students growth were all zero, the new entrants into each educational category would be equal to the number of students enrolled divided by the duration of each category.

[4] For instance, in 1960 Ireland's population is given an average of 14 years of schooling. Considering that most studies (including NSD's) assign less than 14 years to every country for which the years of schooling have been computed in 1990, this figure must be an error.

[5] We use some recent evidence by Lleras-Muney (2005) on education and health in the United States in order to gauge the error induced by ignoring mortality heterogeneity within age groups. We incorporate the largest estimate reported by Lleras-Muney—i.e. a reduction in the unconditional probability of dying of 3.7 percentage points for each additional year of schooling—to correct the number of years of schooling in the US. More specifically, we make an extrapolation for the year 2000 from a 1991 OECD survey. We find that when mortality heterogeneity within age groups is taken into account, the average years of schooling for population 25 and above increases by about 2%. This is a relatively small error.

in mortality rates between age groups. This has strong implications for the accuracy of the estimates, as shown below.

A second problem refers to migration. Our approach implies that immigrants have the same educational level as the population in the host country. But if the host country's population is on average more educated than immigrants, the backward procedure will result in an underestimate of the educational level of host countries (because recent immigration is included in the censuses, but is not, by definition, part of the population in earlier decades). For the same reasons, post-census immigration introduces a positive bias when the forward extrapolation is used to estimate the years of schooling.

In some countries the information available from censuses does not specify the completion rates for some or all the educational categories. In these cases, we assign the completion rates observed in that country at a different date. If this is still not feasible, as in primary and higher education in some high-income OECD countries, we assume full completion. This produces a positive bias in the estimates, although this is not likely to be important since the primary education dropout rates in most of these countries are close to zero. On the other hand, the share of the population with higher education is generally modest (around 15% on average in 1990 and 20% in 2000, for population 25 and above) and the dropout rates are also low. We assign half the number of years of the full duration to an incomplete category.

The calculation for each country is individual in that it uses a different collection of data and relies more or less heavily on assumptions (depending on the number of censuses used). The estimates for some countries (particularly in sub-Saharan Africa) rely entirely on enrolment data. At http://www.iae-csic.uab.es/soto/data.htm we present the basic information used for the calculations: the dates and numbers of censuses on educational attainment for each country, as well as the duration of each educational category. The data on population and enrolment is publicly available from the sources cited above. The same address provides the estimated educational attainment for population aged 15 and above and 25 and above, and the resulting number of years of schooling.

2.2 Data overview

The data set consists of 95 countries, distributed into major regional or economic groups as reported in Table 1. The groups correspond to the Middle East and North Africa (MENA, 8 countries), Sub-Saharan Africa (SSA, 26), Latin America and Caribbean (LAC, 23), East Asia and the Pacific (EAP, 8), South Asia (SA, 3), Eastern Europe and Central Asia (ECA, 4) and High-Income countries[6] (HI, 23). The data was computed for the beginning of each decade from 1960 to 2000, and is accompanied by a projection for 2010. This projection is performed using the forward extrapolation described earlier and is based on population projections by age from the U. S. Census Bureau web site and the estimates of educational attainment for the year 2000.

In 2000, the labour force in high-income countries had an average of 12 years of schooling, while the remaining countries had an average of only 5.7 years. Note the contrast between the relative difference between rich and poor countries and the absolute difference. In relative terms, we observe convergence during the 1960–1990 period, as the ratios have shifted from four to one to two to one. In absolute terms,

[6] In the case of Germany the years of schooling presented here correspond to the unified country.

Table 1 Years of schooling (population 15–64; population-weighted averages)

	1960	1970	1980	1990	2000	2010
All (95)	3.8	4.6	5.3	6.0	6.8	7.4
High-Income (23)	8.7	9.8	10.9	11.6	12.1	12.5
Middle and low income (72)	2.1	2.9	3.7	4.8	5.7	6.5
Middle East & North Africa	0.9	1.6	2.7	4.3	5.9	6.9
Sub-Saharan Africa	1.3	1.7	2.1	3.0	3.9	4.3
Latin America & Caribbean	3.8	4.5	5.3	6.7	7.6	8.2
East Asia & Pacific	2.3	3.2	4.3	5.4	6.4	7.3
South Asia	1.2	1.9	2.6	3.1	4.3	5.3
Eastern Europe & Central Asia	5.3	5.8	6.5	7.1	7.8	8.4

however, the picture is totally different: the gap between rich and poor countries stays essentially constant over the decades (around 6.5 years). According to this criterion, almost no catch-up in schooling has taken place.

Among developing countries, the MENA region displays the highest increase in schooling since 1960. This region has also the fastest growth rate in years of schooling, with an annual rise of 4.8%. The percentage change in years of schooling is also relatively high in the SSA region: it occupies the third place among the most dynamic regions in the world according to this measure. But, if the absolute increase is considered instead, SSA exhibits the lowest improvement. This result stresses the importance of specifying what the proper proxy for human capital in growth regressions should be (the logarithm of the number of years of schooling or its level) and confirms Temple (2001) and Soto (2002) findings.

By 2010, high-income countries will have twelve and a half years of schooling, followed at some distance by ECA countries (the most educated region among developing countries) with only 8.4 years. As a matter of fact, every region in the developing world will have less years of study than the average for high-income countries in 1960. Moreover, by 2010 SSA will be just as educated as LAC was in 1970. Summing up, since the 1960s, and most probably before, Sub-Saharan African countries have exhibited one of the least educated labour forces in the world and there are no signs that this trend will start to be reversed in the coming years.

2.3 Accuracy check

As mentioned above our methodology relies on a number of assumptions. In particular we neglect the fact that two individuals that live in the same country and are of the same age but that differ in their educational levels have plausibly different probabilities of being alive in ten years' time. This omission can lead to a negative bias in the forward extrapolation of years of schooling because more educated individuals are more likely to survive. For the same reason the backward extrapolation may be biased upwards. Another source of estimation error in our methodology is introduced by immigration. When immigration is not taken into account the forward extrapolation leads to an overestimation of years of schooling if immigrant workers are less educated than domestic ones.

In order to check the accuracy of our methodology we compare the estimation of years of schooling obtained from an extrapolation with the years of schooling directly observed from censuses. There are 34 episodes in which a country has censuses for

Table 2 Comparison of 10-year extrapolations with census observations (Years of schooling of population aged 25 and above)

	Year t	Year t + 10
Average census observation	7.72	8.61
Average extrapolation	7.70	8.64
Mean error	−0.02	0.03
Mean absolute value error	0.39	0.38

Source: own calculations based on 34 observations. Extrapolation of year t is obtained from a census in year t + 10. Extrapolation of year t + 10 is obtained from a census in year t

two consecutive decades. These allow us to compare the years of schooling directly measured from a census with the result obtained using a forward extrapolation based on a census made one decade earlier. Similarly we can compare the observed years of schooling with the number that would have been obtained with a backward extrapolation from a census made one decade later. Table 2 summarizes the main results.

The table shows that there are no significant biases either in the backward or forward extrapolation. In both cases the mean error is lower than 1% of the directly observed number of years of schooling. It is possible that the mean error hides some large negative and positive errors that cancel each other out. Still, the mean absolute value error remains reasonably low (about 5%). These preliminary results suggest that our methodology produces realistic estimates of years of schooling. Moreover, ignoring mortality heterogeneity within an age group does not seem to introduce systematic biases.

3 Comparison with other sources

This section compares our data on schooling to the data reported by Barro and Lee (BL, 2001) and De la Fuente and Domenech (2002). The comparison with BL's data is of particular interest since most of the empirical studies of education and growth use their data set as a primary source. Based on UNESCO's database of educational attainment BL have built an upgraded data set for the population aged 25 years and over that attained some level of education.[7] In years when censuses or surveys are not available, BL estimate the educational attainment using enrolment rates according to a perpetual inventory method. Table 3 presents summary measures of years of schooling for the population 25 and above for the 82 countries common to the samples of BL and this paper. Note that in this and all subsequent tables we focus on the years 1960–1990, as opposed to 1960–2000. The literature has studied a shorter period (typically the growth rate between the 1960s and 1990 or earlier). So in order to provide comparable evidence about the quality of our data we decided to stick to that time span.

On average, our series display higher numbers of years of schooling than BL, especially in the case of high-income countries. More importantly, the change over the period 1960–1990 is also higher in our database.

[7] Barro and Lee first published their data in 1993 but we use their latest version for the comparisons presented here since it has arguably less measurement error than the original one.

Table 3 Years of Schooling (population 25 and above; simple averages)

	Barro–Lee			This paper		
	1960	1990	Change	1960	1990	Change
All (82 countries)	3.5 (2.6)	5.6 (2.9)	2.1 (1.1)	3.8 (2.7)	6.2 (3.2)	2.5 (1.1)
High-Income (23)	6.4 (2.1)	8.7 (1.9)	2.3 (1.1)	7.0 (1.9)	9.9 (2.2)	2.9 (0.5)
Middle and low income (59)	2.3 (1.7)	4.4 (2.2)	2.1 (1.0)	2.5 (1.8)	4.8 (2.4)	2.3 (1.2)

Standard deviation in parentheses
Source: authors' calculations based on Barro and Lee (2001) and own data

Although our approach to building the data seems similar to BL's, there are two substantial differences. One concerns the methodology itself and the other, the sources used. On the methodological side, our series are less prone to measurement error than BL series because we make full use of the information by age, whereas BL make only limited use of that information. To illustrate this point, redefine ys_t as the years of schooling of the population aged 25 and above, and let ys^3_{t+5} be the years of schooling of population aged 25–29 in year $t+5$ extrapolated from a census of year t. Let l^3_t be the share of the population aged 25–29 in the total population aged 25 and above, and δ_t the mortality rate of the population aged 25 and above. If ys_t is directly observed from a census, BL's calculation of the years of schooling at time $t+5$ can be expressed as:[8]

$$ys_{t+5} = l^3_{t+5} ys^3_{t+5} + (1-\delta_t)/(1+n_t) ys_t \tag{4}$$

where $n_t = L_{t+5}/L_t - 1$ and L_t is the population aged 25 and above. Equation (4) shows that ys_{t+5} is equal to ys_t adjusted by mortality and population growth, plus the contribution to schooling by the new generation. The change in ys_t between dates t and $t+5$ is given by:

$$ys_{t+5} - ys_t = l^3_{t+5} ys^3_{t+5} - (\delta_t + n_t)/(1+n_t) ys_t \tag{5}$$

Equation (5) highlights a major shortcoming in BL's methodology: it does not account for mortality rate heterogeneity across age groups, whereas older age groups have a higher death rate. Since in most (if not all) countries the younger generations are more educated than older ones the forward extrapolation underestimates the change in years of schooling if death rate heterogeneity is not properly accounted for. Moreover, if the data for a subsequent period is constructed from a census, the downwards bias affecting ys_{t+5} will be reversed. Therefore the change in schooling between $t+5$ and a subsequent date will be biased upwards. Figure 1 illustrates this. At dates t_0 and t_2, the years of schooling are directly measured from censuses, but at t_1 they are estimated with a negative bias. It can readily be shown that under some conditions the correlation between the actual and the estimated change in years of schooling is equal to -1.[9] However, if we were to use a longer time scale, i.e. the change between t_0 and t_2, the bias would disappear. A similar remark is also made by Krueger and Lindahl (2001).

[8] See Barro and Lee (1993), pp 374–375.

[9] The conditions are that the actual and estimated changes in the first period are respectively higher and lower than in the second, as assumed in Fig. 1.

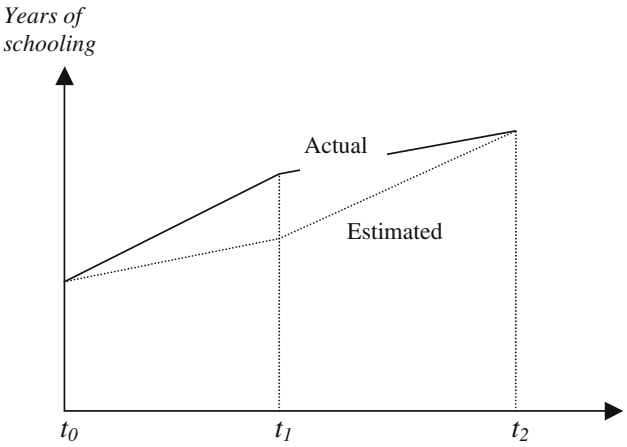

Fig. 1 Measurement error in schooling when mortality heterogeneity is ignored

Unlike BL, we explicitly take into account the heterogeneity in mortality across different age groups by considering the age structure of the population. Slightly modifying (1) and (3) in order to measure the years of schooling of population 25 and above (to maintain coherence with BL's series) we obtain,

$$ys_t = \sum\nolimits_{g=3}^{G} l_t^g ys_t^g \tag{1'}$$

where l_t^g is now defined as the population share of group g in population 25 and above. The forward extrapolation yields,

$$ys_{t+5} = l_{t+5}^3 ys_{t+5}^3 + \sum\nolimits_{g=4}^{G} l_{t+5}^g ys_t^{g-1} \tag{3'}$$

Subtracting (1′) from (3′), we get:

$$ys_{t+5} - ys_t = l_{t+5}^3 ys_{t+5}^3 + \sum\nolimits_{g=4}^{G} \left(l_{t+5}^g - l_t^{g-1}\right) ys_t^{g-1} - l_t^G ys_t^G \tag{6}$$

Defining the mortality rate of group g as $\delta_t^g = \left(L_{t+5}^{g+1} - L_t^g\right)/L_t^g$ where L_t^g is the population of group g, Eq. (6) becomes:

$$ys_{t+5} - ys_t = l_{t+5}^3 ys_{t+5}^3 - \sum\nolimits_{g=4}^{G} \left(\delta_t^{g-1} + n_t\right)/(1+n_t) l_t^{g-1} ys_t^{g-1} - l_t^G ys_t^G \tag{7}$$

Equations (5) and (7) are conceptually the same, except for the fact that in (7) the heterogeneity in mortality rates across age groups is accounted for explicitly. Although this may seem a minor point, assuming a homogenous mortality rate leads to considerably different results. Venezuela provides a good example of how large the differences in estimation can be. UNESCO provides census information on educational attainment in Venezuela for years 1961 and 1981. Table 4 shows the years of schooling reported by Barro and Lee, which we compare to our results.

In 1960 we get virtually the same number of years of schooling as BL because we use the same 1961 census. However, BL find a much smaller increase between 1960 and 1970 than we do. Then, in the following decade, BL find an increase of 2 years,

Table 4 Years of schooling in Venezuela (Population 25 and over)

	Barro–Lee (2001)		This paper	
	Level	Change	Level	Change
1960	2.5		2.5	
1970	2.9	0.4	4.3	1.8
1980	4.9	2.0	5.5	1.2

Source: authors' calculations based on Barro and Lee (2001) and own data

while we obtain a change of 1.2 years only.[10] So over the whole 1960–1980 period, both series provide a similar estimate of the total change in schooling, but the changes between decades in BL's series are more erratic than ours. This example illustrates how noisy the series can be with the hypothesis of homogenous mortality.

Beyond this methodological point, another source of discrepancy between the BL data and ours is that in some cases we use different censuses. For instance, to our knowledge the last data on educational attainment in Jordan published by UNESCO (which is presumably that used by BL) is from 1961. This means that the figures for 1970 and later in the BL database have been filled in with a perpetual inventory method using the 1961 census as a benchmark. On the other hand, the OECD database includes information on the educational attainment by age in Jordan in 1999. We use this data to estimate educational attainment in 1990 using the backward procedure described above. The information provided by the OECD leads to very different figures than those reported by BL: we find that in 1990 the population aged 25 and above had 8.4 years of schooling while BL find 5.4 years. This is one of the highest differences between both series and raises further questions about the reliability of the series. Incidentally, Jordan's statistical office website publishes educational attainment figures for 1994. This shows that the percentage of the population aged 15 and above with preparatory (i.e. first level of secondary education) or full secondary education is 44.3%. This is very close to our estimates for 1990 (45.5%) and is considerably higher than the figure in the BL data set (13.1%). Moreover, the illiteracy rate reported in the website is 15%. Comparing this figure with the 32.2% "no-schooling" population in BL's data and the 13.1% in our data set suggests that our estimates are much closer to reality. We suspect that the rapid growth in the youth (and educated) population in Jordan over recent decades is the cause of the large underestimate in the BL series (between 1961 and 1994 the annual growth of the population aged 15–34 was 5.4%, compared to the 4.5% increase of the population 35 and above).

Finally, it is worth noting some implausible results in BL's database. One example is the data for Austria in 1960, where the sum of the percentages of the population 25 and above assigned to each level of education (including no schooling) totals 84%; or Spain in 1990, where the same operation for the population 15 and above equals 103%. Although, these errors may easily be corrected, there are some features of BL's database that raise more concern. De la Fuente and Domenech (DD, 2006) have already noted the strange pattern followed, among others, of the percentage of the population having attained higher education in Canada. According to BL's data,

[10] The discrepancy in schooling for 1980 is due to the differences in the assumed primary and secondary completion rates. In the case of Venezuela we apply the completion ratio observed in 1960. BL make a regression of the completion ratio on its lagged value and regional dummies and use the fitted values as an estimate of missing completion ratios.

Table 5 Correlation of years of schooling between series (Population 25 and over, 21 OECD Countries)

	Barro–Lee	This paper
Levels (84 observations)		
Barro–Lee	–	0.897
De la Fuente–Domenech	0.908	0.938
First differences (63 observations)		
Barro–Lee	–	0.082
De la Fuente–Doménech	0.105	0.472

higher education increases sharply between 1975 and 1980, and in 1985 it falls back to its previous level. As DD point out, this is the result of classification changes rather than the actual pattern of educational achievement. Besides these classification issues, other results in BL's database are clearly at odds with what one would expect. For example, in 1960 Bolivians aged 15 and over were just as educated as the French population; and in 1980 the average Ecuadorian had more years of study than the average Italian. Summing up, these odd results put in evidence the weaknesses of the BL data set on schooling.

DD (2002, 2006) have built a data set for 21 high-income OECD countries, which they claim to be less noisy than the BL series.[11] Correlations between the three data sets in levels and in first differences are presented in Table 5. For reasons of comparability the correlations are calculated for samples containing the 21 OECD countries. It can be seen that although the correlation for the BL series in levels is fairly high (about 90%), it drops to 10% in first differences. Krueger and Lindahl (2001) have already stressed the low correlation between the series of schooling in first differences. They argue that this is a sign that the series of schooling are too contaminated by measurement error and offer poor informational content. Our series in first-differences are highly correlated with the DD series, which may be an indication of higher quality. However, part of this high correlation may be explained by the use of the same OECD censuses to construct the data and so there is a positive correlation between measurement errors in both series of schooling. We still do not consider that the use of a common source can explain a large share of the correlation of series in first differences since we use a fundamentally different method of estimation for years in which there are no censuses. While we use enrolment data when needed, DD avoid enrolment data to fill in missing data.

Figure 2 displays the different estimates of years of schooling for population 25 and above in OECD countries. The graphs show a neat positive relationship between the different series. The positive association holds for all the decades and is stronger with DD's data than with BL's. The graphs highlight another feature: for each of the decades, BL's data has a tendency to exhibit fewer years of schooling and DD's more than our data set. DD indicate that their data is not directly comparable to BL's since their estimates assume full completion for each level of schooling, whereas BL incorporate estimations of completion rates. Hence, DD's years of schooling data are generally biased upwards.[12]

Krueger and Lindahl (2001) compute reliability ratios to check the quality of the data produced by BL and others. If ys_1 and ys_2 represent two noisy measures

[11] Here we analyse the DD (2002) database, which is the same as the one described in DD (2006).

[12] However this upward bias is not likely to be important since the dropout rates in OECD countries are relatively low.

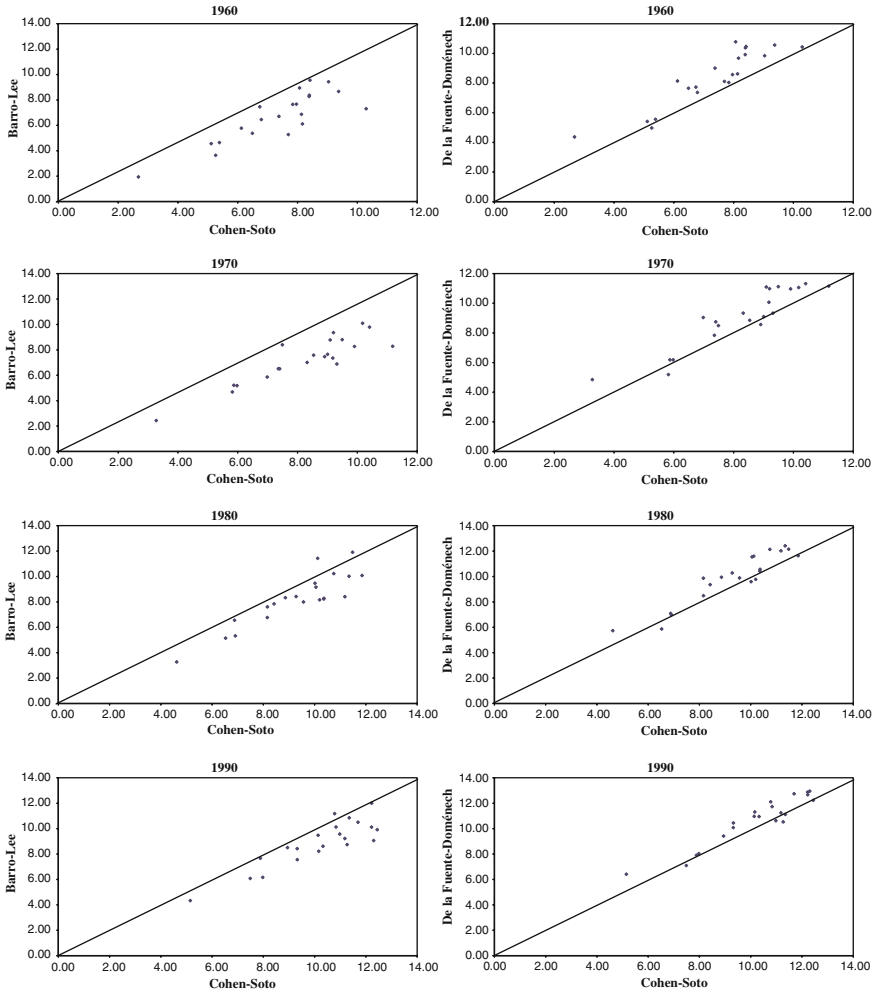

Fig. 2 Comparison of OECD countries

of the true *ys* variable, the reliability ratio of ys_1 with respect to ys_2 is defined as $\mathrm{R}(ys_1, ys_2) = \mathrm{cov}(ys_1, ys_2)/\mathrm{var}(ys_1)$. If the measurement errors of ys_1 and ys_2 are not correlated, the probability limit of $\mathrm{R}(ys_1, ys_2)$ is $\mathrm{var}(ys)/[\mathrm{var}(ys) + \mathrm{var}(e_1)]$ where e_1 is the measurement error of ys_1. Thus, the reliability ratio is the share of the variance of a true variable in the total variance of the variable measured with error. Krueger and Lindahl find that, whereas the reliability ratio is high for the series of schooling in levels, it drops considerably in first-differences. They conclude that the data for first-differences is too noisy to be informative about the actual change in schooling over time.

Table 6 reports the reliability ratios of different measures of the change in years of schooling. In the full sample of countries and 10-year changes our data performs systematically better than BL's data. In the case of the 30-year change over the 1960–90 period the reliability ratios are not statistically different. This is consistent with

Table 6 Reliability of Series in differences (1960–1990)

Period	Countries (Observations)	Reliability of Barro–Lee	Reliability of Cohen–Soto
A. Estimated reliability for all available countries; Barro–Lee and this paper data			
Change, 1960–70	82 (82)	0.33[a] (0.09)	0.40[a] (0.11)
Change, 1970–80	83 (83)	0.22[a] (0.07)	0.43[a] (0.14)
Change, 1980–90	85 (85)	0.39[a] (0.08)	0.59[a] (0.12)
10-year changes, 1960–90	85 (250)	0.35[a] (0.05)	0.55[a] (0.07)
30-year change, 1960–90	82 (82)	0.58[a] (0.10)	0.55[a] (0.09)

Period	Reliability of Barro–Lee		Reliability of De Fuente–Domenech		Reliability of Cohen–Soto	
	(1)	(2)	(3)	(4)	(5)	(6)
B. Estimated reliability for 21 high-income OECD countries; Barro–Lee, De la Fuente–Domenech and this paper data						
Change, 1960–70	.06 (.09)	−.09 (.08)	.36 (.54)	.40[b] (.19)	−.61 (.59)	.49[b] (.23)
Change, 1970–80	−.02 (.07)	−.004 (.06)	−.20 (.73)	.45[a] (.16)	−.06 (.87)	.64[a] (.23)
Change, 1980–90	.01 (.09)	.09 (.07)	.07 (.56)	.26 (.16)	1.04 (.72)	.48 (.29)
10-year changes, 1960–90	.04 (.05)	.03 (.04)	.28 (.34)	.39[a] (.09)	.26 (.40)	.57[a] (.14)
30-year change, 1960–90	−.02 (.12)	−.03 (.10)	−.07 (.44)	.41[b] (.17)	−.13 (.51)	.56[b] (.23)

Notes: Standard error in parentheses. In columns (1) and (6) the benchmark is De la Fuente–Domenech's data. In columns (2) and (4) the benchmark is Cohen–Soto's data. In columns (3) and (5) the benchmark is Barro–Lee's data

[a,b] Significantly different from zero at a 1% and 5% level respectively

Krueger and Lindahl (2001) who argue that for changes over longer periods the variance of the error becomes relatively less important. Second, BL's reliability ratios are not significantly different from zero for OECD countries. These results suggest large measurement errors in the data for these countries in the BL series. Third, DD's data and ours display the highest reliability ratios. As mentioned before, it is possible that a positive correlation between the errors of DD series in first-differences and ours produces upward biases in the reliability ratios. But since DD do not use information on enrolment rates to build their extrapolations for missing years, as we do, it is unlikely that the measurement errors of the series in first-differences are correlated. Overall, the ratios lend some support to the quality of our data. However, these results do not guarantee that our data performs better in growth regressions. This is tested in the following section.

4 Growth and human capital

In this section, we analyze the performance of our data in income growth regressions. The significance of the schooling variable in growth regressions is one additional test for the quality of the data. Indeed, as discussed in the previous section, the low informational content of the schooling series in first differences could partially explain why the earlier literature has found that increases in education are not associated with economic growth. Krueger and Lindahl (2001) and Topel (1999) were the first authors to highlight the lack of informational content of the series in first differences.

The standard equation estimated in the literature is the following:

$$\Delta \log(q_t) = \pi_0 + \pi_1 \Delta \log(k_t) + \pi_2 \Delta \log(h_t) + X_t B + \varepsilon_t \tag{8}$$

where q is output per worker (or per capita), k physical capital per worker, h human capital per worker, X is a set of additional variables that are intended to capture convergence or endogenous growth effects, and ε_t is a residual. Typically X includes the initial levels of income, physical capital or schooling. Some authors also include labour as a third input disembodied from human capital, as in Mankiw, Romer and Weil's model (1992).

Benhabib and Spiegel (1994) first introduced the number of years of schooling as a proxy for human capital, assuming a linear relationship between both variables. This may seem innocuous but (as Sect. 2 illustrated) has strong implications for the measured growth rate of human capital. Pritchett (2001) defines the stock of educational capital as the discounted wage premium of education over raw labour. That is, $h = Cw_0(e^{0.1ys} - 1)$ where C is the discounting factor, w_0 is the wage of labour with no education, and the wage of a worker with ys years of schooling is assumed equal to $w_0 \mathrm{e}^{0.1ys}$. So in Pritchett's regressions, the log of h at date t is given by:

$$Log\,(h_t) = \log\,(C) + \log\,(w_0) + \log\left(e^{0.1ys_t} - 1\right) \tag{9}$$

Note that in Pritchett's formulation the human capital of workers without education is explicitly excluded form the variable h. A broader measure of human capital, which includes human capital of non-educated workers, has gained pre-eminence in empirical macro studies thanks to the works of Bils and Klenow (2000), Heckman and Klenow (1997) and Krueger and Lindahl (2001).[13] This measure is based on the wage regressions estimated by Mincer (1974). In its simplest macroeconomic form, the Mincerian approach implies that the logarithm of human capital is given by:

$$Log\,(h_t) = a + b \times ys_t + e_t \tag{10}$$

where ys_t is the number of years of schooling of the labour force (we ignore the role of experience).

Table 7 summarises the main results of macro regressions reported by different authors. They are all based on income growth from Summers and Heston (1991), different sources for capital stocks, and Barro and Lee (1993) data on years of schooling. Benhabib and Spiegel (1994), who assume a linear relationship between human capital and schooling, find a non-significant and negative coefficient for the log of years of schooling (regression 1). Pritchett's variable for human capital also turns out to have non-significant and negative coefficients (regressions 2 and 3). Krueger and Lindahl (2001), who run the log-change of income on the change in levels of schooling, find positive but non-significant coefficients (regressions 4 and 5). In addition to the disparities in the definitions of the human capital variable, these regressions include different sets of regressors. This renders the comparisons between the different estimates difficult. However, the changes in the human capital variable are systematically non-significant.

The regressions shown in Table 7 are all based on the old Barro and Lee (1993) data on schooling. In Table 8 we replicate the same regressions using the new

[13] This approach has also been adopted by Bloom and Canning (2000), Hall and Jones (1999), Temple (2001) and Topel (1999).

Table 7 Income growth: 1965–1985—Earlier evidence; Dependent Variable: annualized change in log(*GDP*)

	BS	PR		KL	
	(1)	(2)	(3)	(4)	(5)
$\Delta(\log(k))$	.585[a] (.053)	.524[a] (12.8)	.526[a] (12.8)	.598[a] (.062)	.795[a] (.058)
$\Delta(\log(ys))$	−.026 (.071)				
$\Delta(\log(e^{.1\times ys}-1))$		−.049 (1.07)	−.038 (.795)		
$\Delta(ys)$				.066 (.039)	.017 (.032)
ys_{65}				.004[a] (.001)	.0013 (.0008)
$\log(k_{65})$					.016[a] (.002)
$\log(GDP_{65})$	−.166[a] (.030)		.0009 (.625)	−.009[a] (.003)	−.026[a] (.003)
$\Delta(\log(L))$	−.022 (.139)				
R^2	N.A.	N.A.	N.A.	.63	.76
Countries	97	91	91	92	92

Notes: GDP per capita in regressions 1, 4 and 5, and per worker in regressions 2 and 3 (from Summers and Heston, 1991); *k* is capital per worker from different authors; *ys* is years of schooling from Barro and Lee (1993). Standard errors in parenthesis, except for Pritchett who reports t-statistics. Sources: Column 1: Benhabib–Spiegel (1994), Table 2, p. 152; Columns 2 and 3: Pritchett (2001), Table 2, p. 375; Columns 4 and 5: Krueger and Lindahl (2001), Table 5, p. 1125

[a] Significant at a 1% level. Remaining variables are not significant (even at a 10% level)

Barro and Lee (2001) series for the extended 1960–1990 period. Income is GDP per worker from the PWT mark 5.6 and the physical capital data is from Easterly and Levine (2001). To facilitate comparison with Table 7 we use the same numbering for the different regressions so that each column corresponds to the same regression in both tables. Some caution is needed in the comparison of both tables. The GDP and capital data are different and data availability implies that the number of observations in Table 8 is 59. With these caveats in mind Table 8 shows that the new BL data qualitatively provides the same results as before: the change in the schooling variable is not significantly associated with income growth. In most of the new regressions the schooling variable appears with a positive coefficient, except for the regression with Pritchett's definition of human capital that includes initial income (regression 3). But BL's new series fail to show significance in each of the specifications tested. Note however than in Krueger and Lindahl's equation (regression 4), the schooling variable is only marginally non-significant.

Next, we test the performance of our new data set in the same regressions and sample of countries as in Table 8. The results are reported in Table 9, regressions (1) to (5). The schooling variable is not significant in regressions (1) to (3), that is, in regressions run with Benhabib and Spiegel (1994) and Pritchett (2001) specifications for human capital. However, in regressions based on the Mincerian definition (regression 4) our schooling variable is highly significant. Moreover, the point estimate (9.6%) is fairly in line with the average findings of labour studies reported by Psacharopoulos (1994) and Psacharopoulos and Patrinos (2002). The additional introduction of the initial level of physical capital causes the significance of the schooling variables to fall, but the change in schooling is still significant at a 6% level.

The number of countries in regressions (1)–(5) is lower than the one reported in the previous papers. This is because we are analysing the growth rate over the 1960–1990 period, as opposed to the 1965–1985 interval. Since several countries do not have

Table 8 Income growth: 1960–1990 — Barro and Lee (2001) data; Dependent Variable: annualized change in log(*GDP*)

	Regression a la BS	Regressions a la PR		Regressions a la KL	
	(1)	(2)	(3)	(4)	(5)
$\Delta(\log(k))$	.532[a] (.055)	.594[a] (.049)	.595[a] (.047)	.538[a](.053)	.642[a] (.056)
$\Delta(\log(ys))$	.070 (.155)				
$\Delta(\log(e^{.1\times ys}-1))$		.045 (.126)	−.014 (.156)		
$\Delta(ys)$				.061 (.032)	.018 (.023)
ys_{60}				.0016 (.0008)	.0005 (.0006)
$\log(k_{60})$					.010[a] (.003)
$\log(GDP_{60})$	−.0035 (.0021)		−.0019 (.0020)	−.005[b] (.002)	−.016[a] (.004)
$\Delta(\log(L))$	−.437[a] (.137)				
R^2	.74	.70	.70	.73	.79
Countries	59	59	59	59	59

Notes: GDP per worker from Summers and Heston (1991), mark 5.6; *k* is capital per worker from Easterly and Levine (2001). *ys* is years of schooling for population 25 and above from Barro and Lee (2001). Variables in differences are annualized. BS: Benhabib–Spiegel (1994); PR: Pritchett (2001); KL: Krueger and Lindahl (2001). Robust standard errors in parenthesis

[a] Variables are significant at a 1% level

[b] Variables are significant at a 5% level

data for physical capital in 1960, they disappear from the sample. To be sure that our results do not hinge on a reduced number of countries, we replicate the estimates in the shorter 1970–1990 period. This increases the sample size to 81 countries. The estimates that use the Benhabib and Spiegel or Pritchett formulation for human capital (regressions 6 to 8) display non-significant and negative coefficients. In contrast, the estimates based on the Mincerian approach are strongly significant. The coefficient on schooling is now estimated at 12.3% when the initial level of physical capital is omitted (regression 9) and at 9% when it is included (regression 10). Summing up, these regressions show that Benhabib and Spiegel and Pritchett failed to find a significant effect from changes in schooling on growth because their data was too noisy and because they were assuming an inappropriate formulation to represent human capital.

It is important to highlight that these regressions can be criticized for a number of reasons. Namely, several variables that may affect growth are omitted from the regressions and so the estimates presented here may be biased in any direction. Sala-i-Martin (1997) and Sala-i-Martin, Doppelhofer and Miller (2004) find that a considerable number of variables not included in the present regressions are significant and robustly correlated with growth.[14] This may explain why in all regressions the share of physical capital is somewhat high, perhaps as a result of endogeneity bias. However, they are not higher than those reported in the previous papers (see Table 7). Moreover, these regressions offer a good summary of the framework in which the significance of schooling has been evaluated. And the fact is that most of

[14] In contrast Hoover and Perez (2004) and Hendry and Krolzig (2004) find that only a few variables are robustly correlated with growth. In particular they find that different measurements of human capital are not correlated with growth. Similarly, Levine and Renelt (1992) find that only the investment rate and international trade are related to growth.

Table 9 Income growth—This paper's data on schooling; Dependent Variable: annualized change in log(GDP)

	1960–1990					1970–1990				
	Regression a la BS	Regression a la BS		Regression a la KL		Regression a la BS	Regression a la BS		Regression a la KL	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\Delta(\log(k))$	.521[a]	.589[a]	.594[a]	.516[a]	.616[a]	.577[a]	.595[a]	.596[a]	.516[a]	.549[a]
	(.051)	(.047)	(.045)	(.047)	(.058)	(.057)	(.055)	(.055)	(.055)	(.055)
$\Delta(\log(ys))$	.120					−.090				
	(.158)					(.100)				
$\Delta(\log(e^{.1\times ys}-1))$		.068	.018				−.071	−.085		
		(.122)	(.160)				(.082)	(.107)		
$\Delta(ys)$				.096[a]	.049				.123[a]	.090[a]
				(.029)	(.026)				(.025)	(.024)
ys_{60}				.0014[b]	.0005				.0022[a]	.0012
				(.0006)	(.0006)				(.0008)	(.0008)
$\log(k_{60})$					.009[a]					.0082[a]
					(.003)					(.0025)
$\log(GDP_{60})$	−.0034		−.0017	−.0054[b]	−.015[a]	−.0018		−.0005	−.0066[a]	−.015[a]
	(.0021)		(.0020)	(.0021)	(.004)	(.0018)		(.0017)	(.0024)	(.004)
$\Delta(\log(L))$	−.471[a]					−.305				
	(.135)					(.174)				
R^2	.75	.70	.70	.76	.80	.63	.60	.60	.69	.73
Countries	59	59	59	59	59	81	81	81	81	81

Notes: GDP per worker from Summers and Heston (1991), mark 5.6; k is capital per worker from Easterly and Levine (2001); ys is years of schooling for population 25 and above from this paper. Variables in differences are annualized. BS: Benhabib–Spiegel (1994); PR: Pritchett (2001); KL: Krueger and Lindahl (2001). Robust standard errors in parenthesis

[a] Variables are significant at a 1% level

[b] Variables are significant at a 5% level

these regressions have failed to find a significant coefficient associated to the schooling variable. In our view, the findings presented here provide persuasive evidence that our series are less contaminated by measurement error than Barro and Lee's series. Moreover, Krueger and Lindahl state that given the low quality of the earlier series on schooling, they doubt that growth regressions will provide significant coefficients for the schooling variable when physical capital is included in the regressions. The new series presented here do result in significant coefficients even when physical capital is included.

Finally we address the role of outliers as a potential cause of the significance of schooling. Temple (1998) has shown that the presence of some countries may explain the significance of the schooling variable in Mankiw, Romer and Weil's (1992) regressions. Similarly, Temple (1999) provides evidence that by dropping some countries from one of the regressions in Benhabib and Spiegel (1994) the coefficient on schooling increases and becomes significant. We checked the robustness of schooling obtained in regressions 9 and 10 of Table 9 by performing a least trimmed squares estimation as in Temple (1998, 1999). This is aimed at analysing whether the significance of schooling in these regressions is due to the presence of outliers. To proceed we identify a third and a half of the countries with the highest residuals in the sample and then omit them in new regressions. The results are presented in Appendix A.2. There we show that the significance of schooling is not driven by outliers. Indeed, the coefficient for schooling remains remarkably stable in the different samples of countries.

5 Income and human capital: panel estimation

The previous section shows that by using our data on years of schooling to estimate the same equations as the earlier literature we obtain better results than the BL series. However, as mentioned before these regressions have several drawbacks. First, there are endogeneity problems since these are simple OLS regressions. Since there is no use of instrumental variables the estimated coefficients are likely to be biased. And second, these are cross-country regressions and so they do not exploit the time dimension. The few papers that have moved to panel data regressions and that include physical capital have failed to find significance for schooling.

In this section, we estimate a simple production function using recent techniques for panel data regressions. Our aim is simply to analyse whether we can obtain better results with our series than with the BL series. The starting point is an augmented Solow production function for country i at time t as follows:

$$Q_{it} = A_{it} K_{it}^{\alpha} H_{it}^{1-\alpha} \tag{11}$$

where Q_{it} is aggregate income, A_{it} is total factor productivity, K_{it} is aggregate physical capital and $H_{it} = h_{it} L_{it}$. As before h_{it} is human capital per worker and L_{it} is total labour force. Dividing by L_{it} and applying logarithms we obtain:

$$\log(q_{it}) = \log(A_{it}) + \alpha \log(k_{it}) + (1-\alpha) \log(h_{it}) \tag{12}$$

where q_{it} and k_{it} are respectively income and physical capital per worker. In the development accounting literature output is typically expressed in terms of the capital-output ratio (Hall and Jones, 1999). Similarly, Topel (1999) and Soto (2002) express

the production function in terms of the capital-output ratio in order to reduce collinearity problems in estimation. We follow the earlier literature and write (12) as:

$$\log(q_{it}) = \frac{1}{(1-\alpha)}\log(A_{it}) + \frac{\alpha}{(1-\alpha)}\log(k_{it}/q_{it}) + \log(h_{it}) \tag{13}$$

We assume that $\log(h_{it})$ is well represented by the Mincerian approach (10). Finally, total factor productivity is represented as the sum of a fixed effect, a time dummy (which supposes common expected technological growth across countries) and a time varying residual. This leads us to estimate the following equation:

$$\log(q_{it}) = \pi_1 \log(k_{it}/q_{it}) + \pi_2 ys_{it} + \eta_i + \tau_t + \varepsilon_{it} \tag{14}$$

where η_i and τ_t are respectively country and time specific effects and ε_{it} is a residual term. We are able to build an unbalanced panel of 73 countries over the 1960–1990 period. This is the number of countries that appears both in the BL database and our own, and that have physical capital data. As before, the income per worker variable is from Summers and Heston (WPT Mark 5.6). The variables correspond to the beginning of each decade. We first estimate a simple fixed-effect regression as a benchmark against the estimation of instrumental variables. We know that such regression will produce biased estimates for different reasons. First, the income variable q_{it} appears in the capital-output ratio. This will cause a negative bias in the estimated π_1. For the same reason, and since the capital-output ratio is positively correlated with years of schooling, the presence of k/q among the regressors will cause a positive bias on π_2. A second source of bias in fixed-effect estimation is the endogeneity in physical and human capital. If richer countries invest more in both kinds of capital then ignoring this endogeneity will cause positive biases in both π_1 and π_2. However the bias in π_1 introduced by the endogeneity of physical capital is unlikely to be larger than the negative bias described above, i.e. the very presence of q in the capital-output ratio. Finally, measurement errors in both k/q and ys cause negative biases in both coefficients. In all a simple fixed-effect estimation should produce a negatively biased π_1 and an indeterminately biased π_2. In spite of all these drawbacks a fixed-effect estimation is useful as a benchmark estimation against more sophisticated techniques.

Table 10 reports the results. As expected the fixed-effect estimation produces downwardly biased coefficients for the capital-output ratio (regressions 1 and 2). The bias is so large that the coefficients are not significant in both regressions. On the other hand, both schooling variables are significant. As mentioned earlier these coefficients may be biased in any direction. However, the important result for us is that our series display a coefficient that is almost twice as large as the coefficient for the BL series. We interpret this as an indication that our series are less affected by measurement error than the BL series. Indeed, if the only negative bias in this coefficient is caused by measurement error, then the larger the relative weight of the error in the schooling variable, the larger the attenuation bias. Next we check whether this result holds with instrumental variables estimation.

Regressions (3) to (6) show the results obtained with the Blundell and Bond (1998) system GMM estimator. As is well known this is a joint estimation of the equation in levels and in first-differences. For the equations in levels the lagged first-differences of the explanatory variables are used as instruments. For the equations in first-differences the levels of the regressors lagged two or more periods are used as instruments. Blundell and Bond (1998) show that this estimator converges more quickly than

Table 10 Income level—panel estimation; Dependent Variable: Log(PIB per worker), beginning of decade (73 countries; 278 observations, sample period: 1960–1990)

	Fixed-effect Cohen–Soto data (1)	Fixed-effect Barro–Lee data (2)	GMM Cohen–Soto data (3)	GMM Barro–Lee data data (4)	GMM Cohen–Soto (5)	GMM Barro–Lee data (6)
Capital-output ratio	.032 (.103)	.040 (.110)	.680 (.350)	.945[a] (.357)	.700[b] (.336)	.953[a] (.345)
Years of schooling	0.221[a] (.035)	.120[a] (.037)	.126[b] (.053)	.106 (.063)	.123[b] (.051)	.105 (.063)
Second order serial correlation (p-value)			.455	.579	.460	.581
Sargan (p-value)			.143	.282	.250	.379

Notes: Time variables included but not reported; robust standard errors in parentheses. GDP per worker from Summers and Heston (1991), mark 5.6; capital per worker from Easterly and Levine (2001). Instruments in GMM estimation in addition to time dummies: first-differences of explanatory variables lagged one period for the equation in levels; levels of explanatory variables with two lags (regressions 3 and 4) and two and three lags (regressions 5 and 6) for equations in first-differences

[a] Variables are significant at a 1% level

[b] Variables are significant at a 5% level

the Arellano and Bond (1991) estimator when the explanatory variables are highly autocorrelated, which is the case here. Regarding the choice of lags to be used as instruments in the equation in first differences, we use the levels of explanatory variables with two lags (regressions (3) and (4)) and two and three lags (regressions (5) and (6)). Our series of schooling is significant at a 5% level in both estimations (regressions (3) and (5)). By contrast, the BL variable is never significant (regressions (4) and (6)). Note that the coefficients are smaller than in fixed-effect estimation, which suggests that in these regressions the endogeneity bias is more important than the attenuation bias.

Not surprisingly, the capital-output ratio displays considerably larger coefficients than the ones obtained by fixed-effect estimation. The coefficient is marginally not significant in regression (3) but it is significant in the remaining regressions. In the equations estimated with our series the implicit physical capital share is about 40%, whereas in the regressions with the BL data it is about 49%. Finally, in every regression the hypothesis that the instruments are exogenous is not rejected, as the second order serial correlation and Sargan tests show.

From regressions (3) and (5) the long-term impact on income of one additional year of schooling is slightly above 12%. This is close to the typical Mincerian return found in labour studies. However, rather than stressing the magnitude of the coefficient for the schooling variable, we consider that the main result from these regressions is that with our series we can obtain significant coefficients, even with GMM estimation.

Interestingly, the difference in the coefficient for schooling between BL data and our own is considerably lower in GMM estimation than in fixed-effect estimation. This is consistent with the idea that instrumentation techniques tend to correct the bias

introduced by measurement error. Our data, however, yields significant coefficients while the BL series do not.

6 Conclusion

In this paper, we present a new data set for years of schooling intended to reduce the measurement error present in existing series. We concur with Krueger and Lindahl (2001) in the debate on errors in the measurement of data on schooling and the implications for the estimated impact of schooling on income. Furthermore, our results confirm those of De la Fuente and Domenech (2002, 2006), who highlight the low quality of schooling data even for the subgroup of high-income OECD countries.

One of the main features of these series is that they are built by taking into account the age structure of the population. In particular we consider the fact that older people, who on average have a lower educational level, have higher mortality rates. This allows us to make more accurate estimates of the educational level of the labour force for years in which we cannot observe it directly from censuses or surveys. Previous series do not account for mortality heterogeneity among age groups to the detriment of data quality. In addition, we have avoided the use of sources that employ different classification systems of education in a country over time. The approach we follow leads to a reduction in measurement error, particularly in the time dimension.

Our data performs better than Barro and Lee (2001) series in the standard cross-country growth regressions estimated in the earlier literature. Indeed when our data is entered into such regressions we find that it is significant, as opposed to the Barro and Lee (2001) series. This holds even for the kind of regressions estimated by Krueger and Lindahl (2001), who failed to find significance for the Barro and Lee (2001) series when physical capital is included among the regressors. Moreover, our series are also significant in more sophisticated panel data regressions that attempt to account for endogeneity or problems identifying the effect of schooling. Indeed when we estimate an augmented Solow production function that embeds the Mincerian approach to human capital we find that our series are highly significant. Moreover, the estimated long-term effect of schooling is close to the typical micro Mincerian return. This result suggests the absence of externalities to education, which is consistent with Acemoglu and Angrist (2001) and Ciccone and Peri (2005) among others. But this should be a matter of further empirical investigation. The main contribution of this paper is to make a new and reliable data set on schooling available for a large group of countries that may prove useful for further research into human capital.

Acknowledgements This work was started as part of the OECD Development Centre research agenda. It does not necessarily represent the views of the OECD. Soto received financial support from the Spanish Ministry of Education and Science under project SEC2002–01612. We would like to thank Sabrina Chastang for her valuable help at an early stage of this paper. The full data set presented here and information about the sources are available at http://www.iae-csic.uab.es/soto/data.htm.

A Appendix

A.1 Estimation of net intake ratios

This appendix describes how enrolment figures are used to estimate net intake ratios and educational attainment. The main problem with estimating intake ratios is accounting for students that have dropped out and thus that do not appear in the enrolment figures. Neglecting dropouts leads to underestimations of the actual intake rates and thus the educational attainment of the population. This is not an important source of bias for most of the OECD countries because dropout rates there are relatively low. But developing countries, and especially low-income countries, display dropout rates as high as 15%. Another source of bias is the presence of repeaters in the enrolment data, which leads to overestimations of the number of students that have received formal education. Although existing measures of educational attainment generally adjust their estimates by the repeater effect, they fail to take into account the dropout effect.

The present procedure estimates net intakes from enrolment figures. Calling N_t the net intakes in year t, d the drop out rate, r the repetition rate and P the duration in years of primary school, $(1 - \mathrm{d} - \mathrm{r})^P \times N_t$ students will succeed in finishing primary school in P years.

Making the reasonable assumption that each student may repeat a maximum of three times in primary education, each grade is composed of students that have never repeated and students that have repeated once, twice or three times. Calling n the growth rate of net intakes, the expression linking primary enrolment E_t to net intakes N_t in year t is:

$$E_t = N_t \sum_{j=0}^{P-1} (1-d-r)^j \left[\frac{r^3 C'(j+1,\,3)}{(1+n)^{j+3}} + \frac{r^2 C'(j+1,\,2)}{(1+n)^{j+2}} + \frac{r C'(j+1,\,1)}{(1+n)^{j+1}} + \frac{C'(j+1,\,0)}{(1+n)^j} \right] \quad \text{(A.1)}$$

where $C'(j+1, i)$ is a combinatorial with repetition of i out of $j+1$ years. We use expression (A.1) to build net intakes series based on enrolment data. The right-hand side of (A.1) is a function of three parameters: the repetition rate (r), the dropout rate (d) and the net intake growth rate (n). In the stationary state, primary enrolment E_t grows at the same rate as net intakes. Thus n may be computed from the enrolment growth rate. One particular case is when $d = r = n = 0$. In this case, $E_t = N_t \times P$ and so the number of new entrants is simply equal to the pupils enrolled in primary education divided by its duration. UNESCO provides estimates for repetition and survival rates in primary schooling for most countries in the world starting in 1970. The survival rate (which is defined as the percentage of students enrolled in the first grade that are expected to reach the final grade) is used to compute the dropout rate. Defining s as the survival rate and noting that,

$$s = (1-r-d)^P (1 + rP + r^2 C'(P,\,2) + r^3 C'(P,\,3))$$

it can be deduced that the dropout rate is equal to,

$$d = (1-r) - \left[\frac{s}{(1 + rP + r^2 C'(P,\,2) + r^3 C'(P,\,3))} \right]^{\frac{1}{P}}$$

In order to estimate the percentage population that has completed primary school we multiply the survival rate by the net intake rate obtained from (A.1). Finally, a similar procedure is used to estimate attainment in secondary and higher education.

Not all the countries have full information on enrolment or on the parameters involved in (A.1). In several cases, especially in African countries, population data is very limited and available only back as far as 1950. In these cases, it is assumed that net intake rates before 1950 were the same as in that year. While this assumption may appear unrealistic, it is unlikely to produce major errors because, as the data shows, enrolment and net intake rates were very low in 1950 and close to zero in secondary and higher education. Thus the error will be limited to (the very low) participation in primary education.

In other cases, like the years around the two world wars, there is no enrolment data for most of European countries. For these cases, the information is taken from the closest year with available data. This procedure is unlikely to lead into large biases, since enrolment figures change little from year to year.

A.2 Least trimmed squares estimation

Here we replicate some of the main regressions reported in Table 9 to analyse their robustness. We are particularly interested in finding out whether the significance of schooling depends on the presence of certain outliers. Temple (1998, 1999) shows that the coefficient for schooling varies considerably under certain specifications when the countries with relatively high residuals are eliminated from the sample. We again estimate regressions (9) and (10) by dropping a third and a half of countries with the highest residuals (the results did not change when a lower number of countries were eliminated). In Table A.2 we observe that omitting these countries from the regressions produces small changes in the coefficients. If anything, the significance of

Table A.2 Least trimmed squares estimation. Income growth: 1970–1990; Dependent Variable: annualized change in log(GDP)

	(1) Full sample	(2)	(3)	(4) Full sample	(5)	(6)
$\Delta(\log(k))$	.516[a]	.525[a]	.525[a]	.549[a]	.510[a]	.530[a]
	(.055)	(.028)	(.026)	(.055)	(.028)	(.019)
$\Delta(ys)$	.123[a]	.121[a]	.126[a]	.090[a]	.122[a]	.105[a]
	(.025)	(.015)	(.012)	(.024)	(.014)	(.011)
ys_{70}	.0022[a]	.0019[a]	.0020[a]	.0012	.0012[a]	.0014[a]
	(.0008)	(.0004)	(.0003)	(.0008)	(.0004)	(.0003)
$\log(k_{70})$				.0082[a]	.0081[a]	.0079[a]
				(.0025)	(.0014)	(.0008)
$\log(GDP_{70})$	−.0066[a]	−.0058[a]	−.0064[a]	−.015[a]	−.016[a]	−.016[a]
	(.0024)	(.0012)	(.0010)	(.004)	(.002)	(.001)
R^2	.69	.93	.96	.73	.94	.97
Countries	81	54	41	81	54	41

Notes: GDP per worker from Summers and Heston (1991), mark 5.6; k is capital per worker from Easterly and Levine (2001); ys is years of schooling for population 25 and above from this paper. Variables in differences are annualized. Robust standard errors in parenthesis. Regressions (2) and (3) omit a third and a half of countries with the highest residuals in regression (1). Regressions (5) and (6) omit a third and a half of countries with the highest residuals in regression (4)

[a] Variables are significant at a 1% level

schooling is even stronger than in the full sample regressions because of the fall in standard errors. We proceed similarly with the other specifications of Table 9 and find the same results as there (we do not report these results). Namely, the regressions a la Benhabib and Spiegel (1994) and Pritchett (2001) do not produce significant coefficients for the schooling variable. Overall, these results suggest that the significance of schooling in these regressions is not caused by unrepresentative observations.

References

Acemoglu, A., & Angrist, J. (2001). How large are human-capital externalities? Evidence from compulsory schooling laws. *NBER Macroeconomics Annual, 2000*, 9–59.

Arellano, M., & Bond, S., (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies, 58*(2), 277–297.

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

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

Benhabib, J., & Spiegel, M. M. (1994). The role of human capital in economic development: Evidence from aggregate cross-country data. *Journal of Monetary Economics, 34*(2), 143–173.

Bils, M., & Klenow, P. (2000). Does schooling cause growth? *American Economic Review, (90)*5, 1160–1183.

Bloom, D., & Canning, D. (2000). Health, human capital and economic growth, commission on macroeconomics and health. mimeo.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics, 87*, 115–143.

Ciccone, A., & Peri, G. (2005). Identifying human capital externalities: Theory and applications. *Forthcoming in the Review of Economic Studies*.

Cohen, D., (1996). Tests of the convergence h ypothesis: Some further results. *Journal of Economic Growth, 1*(3), 351–361.

De la Fuente, A., & Domenech, R. (2002). Educational attainment in the OECD, 1960–1995. *CEPR DP 3390*.

De la Fuente, A., & Domenech, R. (2006). Human capital in growth regression: How much difference does quality data make? *Journal of the European Economic Association, 4*(1), 1–36. An earlier version was published in CEPR DP 2466 (2000).

Easterly, W., & Levine, R. (2001). It's not factor accumulation: Stylized facts and growth models. *The World Bank Economic Review 15*(2), 177–219.

Hall, R., & Jones, C. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics, 114*(1), 83–116.

Heckman, J., & Klenow, P. (1997). *Human capital policy*. mimeo, University of Chicago.

Hendry, D., & Krolzig, H. (2004). We ran one regression. *Oxford bulletin of Economics and Statistics, 66*(5), 799–810.

Hoover, K., & Perez, S. (2004). Truth and robustness in cross-country growth regressions. *Oxford bulletin of Economics and Statistics, 66*(5), 765–798.

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

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *American Economic Review, 82*(4), 942–963.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *Review of Economic Studies, 72*, 189–221.

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

Mankiw, G., Romer, D., & Weil, D. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics, 107*(2), 402–437.

Mincer, J. (1974). *Schooling, experience and earnings*. Columbia University Press.

Mitchell, B. R. (1993). *International historical statistics: Africa, Asia and Oceania 1750–1988*. New York, NY: M Stockton Press.

Mitchell, B. R. (1998a). *International historical statistics: Europe 1750–1993*. New York, NY: M Stockton Press.
Mitchell, B. R. (1998b). *International historical statistics: The Americas 1750–1993*. New York, NY: M Stockton Press.
Nehru, V., Swanson, E., & Dubey, A. (1995). A new database on human capital stocks in developing and industrial countries: Sources methodology and results. *Journal of Development Economics, 46*(2), 379–401.
OECD. (2000). *Investing in education: Analysis of the 1999 World education indicators*. Paris.
OECD. (2003). *Financing education – Investment and returns: Analysis of the World Education Indicators*. Paris.
OECD. (various issues). *Education at a glance: OECD indicators*. Paris.
Pritchett, L. (2001). Where has all the education gone?. *World Bank Economic Review, 15*(3), 367–391.
Psacharopoulos, G. (1994). Returns to investment in education: A global update. *World Development, 22*(9), 1325–1343.
Psacharopoulos, G., & Patrinos, H. (2002). Returns to investment in education: A further update. *The World Bank Policy Research Working Paper Series* 2881.
Romer, P. (1990). Endogenous technological change. *Journal of Political Economy, Part 2, 98*(5), S71–102.
Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review, 87*(2), 178–183.
Sala-i-Martin, X., Doppelhofer, G., & Miller, R. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review, 94*(4), 813–835.
Soto, M. (2002). Rediscovering education in growth regressions. OECD Development Centre Technical Paper 202.
Summers, R., & Heston, A. (1991). The Penn World table (Mark 5): An expanded set of international comparisons, 1950–1988. *The Quarterly Journal of Economics, 106*(2), 327–368.
Temple, J. (1998). Robustness tests of the augmented Solow model. *Journal of Applied Econometrics, 13*, 361–375.
Temple, J. (1999). A positive effect of human capital on growth. *Economic Letters, 65*, 131–134.
Temple, J. (2001). Generalizations that aren't? Evidence on education and growth. *European Economic Review, 45*(4–6), 905–918.
Topel, R. (1999). Labor markets and economic growth. In O. Ashenfelter, & D. Card (Eds.), Handbook of labor economics, (Vol. 3, pp. 2943–2984). Amsterdam: Elsevier Science.
UNESCO. (1977). *Statistics of educational attainment and illiteracy: 1945–1974*. Paris.
UNESCO. (1983). *Statistics of educational attainment and illiteracy: 1970–1980*. Paris.
UNESCO. (various issues). *Statistical Yearbook*. Paris.