HUMAN CAPITAL IN GROWTH REGRESSIONS: HOW MUCH DIFFERENCE DOES DATA QUALITY MAKE?

Angel de la Fuente
Instituto de Análisis Económico (CSIC) and CREA

Rafael Doménech
Universidad de Valencia

Abstract
We construct estimates of educational attainment for a sample of 21 OECD countries. Our series incorporate previously unexploited information and remove sharp breaks in the data that can only reflect changes in classification criteria. We then construct indicators of the information content of our estimates and a number of previously available data sets and examine their performance in several growth specifications. We find a clear positive correlation between data quality and the size and significance of human capital coefficients in growth regressions. Using an extension of the classical errors in variables model to correct for measurement error bias, we construct a set of meta-estimates of the coefficient of years of schooling in an aggregate Cobb-Douglas production function. Our results suggest that the value of this parameter is likely to be above 0.60. (JEL: O40, I20, O30, C19)

1. Introduction

Empirical investigations of the contribution of human capital accumulation to economic growth have often produced discouraging results. Educational variables frequently turn out to be insignificant or to have the “wrong” sign in growth regressions, particularly when these are estimated using first-differenced or panel specifications.1 In this paper we provide evidence that these counterintuitive results on human capital and growth can be attributed to deficiencies in the data and show that improvements in data quality lead to larger and more precise estimates of schooling coefficients in growth regressions.

Acknowledgments: This paper is part of a research project cofinanced by the European Fund for Regional Development and Fundación Caixa Galicia. We gratefully acknowledge additional support from the Spanish Ministry of Education and Science through grants SEC2002-01612 and SEC2002-0026 and from the OECD. We would like to thank Georges Lemaitre and Jean-Luc Heller of the statistics and indicators division of the OECD and their correspondents in the national statistical offices of member countries who have contributed greatly to the improvement of this data set. We are also grateful to Jon Temple, Antonio Ciccone, Jong-Wha Lee, José Emilio Boscá, Anna Borkowsky, Gunilla Dahlen, Roberto Perotti, and several anonymous referees for their helpful comments and suggestions; and to María Jesús Freire and Juan Antonio Duro for their competent research assistance.

E-mail addresses: de la Fuente: Angel.delaFuente@uab.es; Doménech: Rafael.Domenech@uv.es.

The paper is organized in the following way. Section 2 documents a number of suspicious features and inconsistencies in the main schooling data sets used in the empirical growth literature. In Section 3, we discuss the construction of new attainment series for a sample of 21 OECD countries and argue that the new data increase the signal-to-noise ratio relative to earlier sources. Our series make use of previously unexploited information, including unpublished data supplied by the OECD and by a number of national statistical agencies, and remove sharp breaks in the data that can only be due to changes in classification criteria.

In Section 4, we systematically compare the performance of eight different schooling series in a number of standard growth specifications. We find that ordinary least squares (OLS) estimates of the coefficient of years of schooling in an aggregate production function become considerably larger and more precise when we use recently constructed attainment series that can be expected to have a greater information content than previous data sets. To provide firmer evidence that this is indeed the case, Section 5 deals with the construction of statistical indicators of data quality. We extend the procedure developed by Krueger and Lindhal (2001) to estimate reliability ratios in a way that allows us to exploit the availability of multiple schooling series to increase the precision of the estimates and to allow measurement error to be correlated across data sets. We find that, for the case of the OECD sample we work with, our series have the highest information content of all those available, followed closely by those constructed by Cohen and Soto (2001). We also find that there is a systematic relationship between data quality and the size of OLS estimates of the coefficient of human capital. In Section 6, we show how this relationship can be exploited to correct for measurement error bias, essentially by extrapolating it to the limiting case of perfectly measured schooling data. The exercise yields a set of consistent meta-estimates of the coefficient of schooling in the aggregate production function that are considerably larger than those found in the previous literature.

2. Cross-country Schooling Data: A Brief Review and Some Troubling Features

Most governments gather information on a number of educational indicators through population censuses, labour force surveys, and specialized studies and surveys. Various international organizations collect these data and compile comparative statistics that provide easily accessible and (supposedly) homogeneous information for a large number of countries. The most comprehensive regular source of international educational statistics is UNESCO’s *Statistical Yearbook*. This publication provides reasonably complete yearly time series on school enrollment rates by level of education for most countries in the world and contains some data on the educational attainment of the adult population, government expenditures on education, teacher/pupil ratios, and other variables of interest.
The UNESCO enrollment series have been used in a large number of empirical studies of the effects of education on aggregate productivity. In many cases this choice reflects the easy availability and broad coverage of these data rather than their theoretical suitability for the purpose of the study. Enrollment rates can probably be considered an acceptable, although imperfect, proxy for the flow of educational investment. On the other hand, this variable is not necessarily a good indicator of the existing stock of human capital because average educational attainment (which is often the more interesting variable from a theoretical point of view) responds to investment flows only gradually and with a very considerable lag.

In an attempt to remedy these shortcomings, a number of researchers have constructed data sets that attempt to measure directly the educational stock embodied in the population or labour force of large samples of countries during a period of several decades. The best known attempts in this direction are the work of Kyriacou (1991), the different versions of the Barro and Lee data set (1993, 1996, 2000), and the series constructed by Lau, Jamison, and Louat (1991), Lau, Bhalla, and Louat (1991) and Nehru, Swanson, and Dubey (1995). (We abbreviated these as follows: Kyriacou (1991) as KYR; Barro and Lee (1993, 1996, 2000) collectively as B&L and individually as B&L93, B&L96, and B&L00; Nehru, Swanson, and Dubey (1995) as NSD.)

We can divide these studies into two groups according to whether they make use of both census attainment data and enrollment series or only the latter to obtain series of average years of schooling and the educational composition of the population. The first set of papers (KYR and B&L) relies on census figures where available and uses enrollment data to fill in the missing values. KYR uses a simple regression of educational stocks on lagged flows to estimate the unavailable levels of schooling. This procedure is valid only if the relationship between these two variables is stable over time and across countries, an assumption that seems unlikely a priori and is not tested. B&L, by contrast, use a combination of interpolation between available census observations (where possible) and a perpetual inventory procedure that makes use of enrollment data to estimate changes from nearby (either forward or backward) benchmark observations. In principle, this approach should be superior to KYR’s because it makes use of more information and does not rely on such strong implicit assumptions.

The second group of papers (Lau, Jamison, and Louat and NSD) uses only enrollment data to construct time series of educational attainment. The version of the perpetual inventory method used in these studies takes into account more factors than the one used in the first version of B&L, particularly in the case of NSD.2

2. Differences across these studies have to do with the correction of enrollment rates for dropouts and repeaters and with the estimation of survival probabilities. Latter versions of Barro and Lee have improved the treatment of the first of these issues.
On the other hand, these studies completely ignore census data on attainment levels. To justify this decision, NSD observe that census publications typically do not report the actual years of schooling of individuals (only whether or not they have started or completed a certain level of education) and often provide information only for the population aged 25 and over. As a result, there will be some arbitrariness in estimates of average years of schooling based on these data, and the omission of the younger segments of the population may bias the results, particularly in less developed countries (LDCs), where this age group is typically very large and much more educated than older cohorts. Although this may call for adjustment of the census figures on the basis of other sources, in our opinion it does not justify discarding the only direct information available on the variables of interest.

Methodological differences in the construction of different schooling series would be of relatively little concern if they all gave us a consistent and reasonable picture of educational attainment levels across countries and of their evolution over time. As we will see presently, this is not the case. The analysis of the different series reveals very significant discrepancies, among them the relative positions of many countries and implausible estimates or time profiles for several of them. Although the various studies generally coincide when comparisons are made across broad regions (e.g., the OECD vs. LDCs in various geographical areas), the discrepancies are very important when we focus on the group of industrialized economies. Another cause for concern is that existing estimates often display extremely large changes in attainment levels over periods as short as five years (particularly at the secondary and tertiary levels). Many of the suspicious features of the data we will document seem to reflect deficiencies in the primary statistics used to construct them, which do not appear to be consistent, across countries or over time, in their treatment of vocational and technical training and other courses of study, and reflect at times the number of people who have started a certain level of education and, at others, those who have completed it.3

To illustrate these problems and to get some feeling for the overall reasonableness of the existing data, in the remainder of this section we will take a closer look at the B&L96 and NSD data sets. These are the most sophisticated data sets within each of the groups identified above that were available in the mid-1990s, and they have been used in many growth studies. As in the rest of the paper, we will concentrate on a sample of OECD countries. One of the main reasons for this choice is that educational statistics for this set of advanced industrial nations

---

3. Steedman (1996) documents the existence of important inconsistencies in the way educational data are collected in different countries and argues that this problem can significantly distort the measurement of educational levels. She notes, for example, that countries differ in the extent to which they report qualifications not issued directly (or at least recognized) by the state and that practices differ as to the classification of courses which may be considered borderline between different International Standard Classification of Education (ISCED) levels. Behrman and Rosenzweig (1994) also discuss some of the shortcomings of UNESCO’s educational data.
Figure 1. Average years of schooling in 1985: B&L96 vs. NSD.
Notes: The estimates refer to the population over 15 in the case of B&L96 and to the age group 15–64 in NSD.
The estimated equation is of the form $H_{NSD} = 4.50 + 0.503 H_{B&L}$, $t = 3.21$, $R^2 = 0.329$. The flattest line in the figure is the regression line fitted after excluding the four countries with the lowest schooling levels. The thinnest and steepest line is the “diagonal,” where all the observations would fall if both sources agreed.

Aus = Australia; Be = Belgium; Can = Canada; CH = Switzerland; Dk = Denmark; Fin = Finland; Fr = France; Ge = West Germany; Gr = Greece; Is = Iceland; Ir = Ireland; It = Italy; Jap = Japan; Nl = Netherlands; Nor = Norway; NZ = New Zealand; Ost = Austria; Po = Portugal; Sp = Spain; Swe = Sweden; Tu = Turkey; UK = United Kingdom; USA = United States.

are presumably of decent quality. Any deficiencies we find in them are likely to be compounded in the case of poorer countries.

The degree of consistency between these two sources varies a great deal depending on the level of aggregation we consider. The overall correlation (computed over common observations) between B&L’s and NSD’s estimates of average years of schooling is 0.841. An examination of average values for different geographical regions and of their evolution over time also reveals a fairly consistent and reasonable pattern. Industrialized countries and socialist economies display much higher attainment rates than LDCs. Within this last group, Africa lies at the bottom of the distribution, whereas Latin America does fairly well and Southeast Asia features the largest improvement over the period.

This high overall correlation, however, hides significant discrepancies between the two data sets, both over time and across countries. Figure 1 shows B&L96’s and NSD’s estimates of the average years of total schooling of the population aged 15 and over for OECD countries in 1985. The correlation for the 23 countries in this sample (there are no data for Luxembourg) is now 0.574, but drops to zero (0.063) when we exclude the four countries with the lowest levels
Figure 2. Average years of total and primary schooling in the OECD: B&L96 vs. NSD.  
Notes: Unweighted averages over the available OECD countries. Neither source reports data for Luxembourg. The sample excludes New Zealand except for average total years of schooling, because NSD only provide data on this variable but not its breakdown by level. The data are for the age group 15–64 in the case of NSD and for the population aged 15 and over in B&L96. The last year for the NSD series is 1987, rather than 1990.

of schooling. When we disaggregate, the correlation is fairly high at the university level (0.767) and much lower for primary (0.362) and secondary (0.397) attainment.

There are also significant differences between the B&L and NSD series in terms of their time profiles. This is illustrated in Figure 2, which shows the evolution of average years of total and primary schooling ($H$ and $H1$) in the average OECD country. In terms of years of total schooling, both data sets display an increasing trend, but the rise in average attainment is much more marked in the case of B&L. At the primary level, NSD’s attainment figures are implausibly high and display an extremely suspicious downward trend.

Another disturbing feature of the human capital series is the existence of sharp breaks and implausible changes in attainment levels over very short periods. This problem affects the B&L data set much more than the NSD series, which are much smoother by construction. As an example, Figure 3 shows the evolution of the B&L96 university attainment rates for the population over 25 in a number of countries that display extremely suspicious patterns. In all cases, the sharp break in the series signals in all probability a change of criterion in the elaboration of educational statistics. In the case of Canada, for instance, the detailed data provided by Statistics Canada indicates that the sudden drop in university attainment between 1980 and 1985 found in the B&L series can be explained by the exclusion of postsecondary vocational qualifications in the latter year.
A similar pattern of breaks in the schooling series can be found in a large number of countries, making for an implausibly large range of values of the annual growth rate of years of schooling. Restricting ourselves to an OECD sample of 21 countries, B&L’s estimates of this variable range between $-1.35\%$ and $7.80\%$; moreover, 15.9% of their observations are negative, and 19% of them exceed 2%. Figure 4 shows that this is a common feature of all versions of the B&L data set. We believe that this anomaly, which seems to arise from these authors’ reliance on UNESCO data, cannot be corrected by any improvements in the fill-in procedure alone. We also suspect that this may explain why these series often generate implausible results in growth regressions, particularly when they are estimated using panel or differenced specifications that rely heavily on the time variation of the data.

The preceding discussion is far from providing an exhaustive list of the suspect features of different educational data sets. On the other hand, it is probably

![Figure 3. Evolution of university attainment levels: Australia, New Zealand, and Canada. Source: B&L96. Population aged 25 and over.](image)

![Figure 4. Fitted distribution of the annual growth rate of years of schooling in an OECD sample, different versions of the B&L data set.](image)
sufficient to conclude that there are good reasons to be concerned about the quality of the educational data used in the literature and about the implications of measurement error for empirical estimates of the growth effects of investment in education. Such concerns have motivated some recent studies that attempt to increase the signal-to-noise ratio in the schooling series by exploiting additional sources of information and introducing various corrections. The latest version of the Barro and Lee data set (B&L00) incorporates various refinements of the procedure used to fill in missing observations. De la Fuente and Doménech (2000) and Cohen and Soto (2001) (abbreviated D&D00 and C&S, respectively), on the other hand, have focused on cleaning up the available census and survey data. In the following section we will discuss the construction of attainment series that update the work in the first of these papers.


As we have seen in the previous section, the available schooling data contain a large amount of noise that can be traced back to inconsistencies in the underlying primary statistics. To reduce this noise, we have constructed educational attainment series for the adult population of a sample of 21 OECD countries covering the period between 1960 and either 1990 or 1995. The complete series and a detailed set of country notes can be found in the Data Appendix to the present paper (available at ⟨http://iei.uv.es/~rdomenec/human/human.html⟩) or in de la Fuente and Doménech (2002), (abbreviated D&D02).

The data set described in this section is a revised and partially extended version of the one constructed in D&D00. The most important change relative to our earlier estimates is that we have incorporated unpublished information supplied by the OECD and by the national statistical offices of around a dozen member states in response to a request for assistance that was channeled through the statistics and indicators division of the OECD. This organization sent our previous paper (D&D00) to their education correspondents in the statistical offices of member countries and asked them to check our series and to provide any additional information that may help improve them. Their responses have been extremely helpful in a number of cases where the published information was quite limited.

We aim to provide estimates of the fraction of the population aged 25 and over that has started (but not necessarily completed) each of the following levels of education: primary schooling ($L_1$), lower and upper secondary schooling ($L_2.1$ and $L_2.2$), and two levels of higher education ($L_3.1$ and $L_3.2$). For some

---

4. Iceland, Luxembourg, Turkey, and recent OECD members are left out because of the paucity of information.
countries, however, the available data may refer to a different age group or to those who have completed each schooling level, and it is not always possible to detect when this is the case. We have tried to include advanced vocational training in the first level of higher attainment. We also report illiterates (L0) as a separate category in the four countries where illiteracy rates are significant during the sample period (Portugal, Greece, Spain, and Italy). For the rest of the sample, the lowest reported category is L1, and it includes all those who have not reached secondary school.

To construct our schooling series we first collected all the information we could find on the distribution of the adult population by educational level in 21 OECD countries. We used both international publications and national sources (census reports and surveys, statistical yearbooks, and unpublished national and OECD data). Next, we tried to construct a plausible time profile of attainment in each country using all the available data. For those countries for which reasonably complete series are available, we have relied primarily on national sources. For the rest, we start from the most plausible set of attainment estimates available around 1990 or 1995 and proceed backwards, trying to avoid jumps in the series that can only reflect changes in classification criteria. The construction of the series often involved subjective judgments to choose among alternative census or survey estimates when several are available. In a few cases, we have also reinterpreted some of the data from international compilations as referring to somewhat broader or narrower schooling categories than the reported one. Missing data points lying between available census observations are filled in by simple linear interpolation. We have avoided the use of flow estimates based on enrollment data because they seem to produce implausible time profiles. Missing observations prior to the first available census are estimated, whenever possible, by backward extrapolations that make use of census information on attainment levels disaggregated by age group.

Data availability varies widely across countries. Table 1 shows the fraction of the reported data points that correspond to “direct observations” (taken from census or survey reports) and the earliest and latest such observations available for secondary and higher attainment levels. The number of possible observations is typically either 21 or 24 for each level of schooling depending on whether the series ends in 1990 or 1995 (two sublevels and a total times seven or eight quinquennial observations). In the case of Italy, there seem to be no short higher education courses, so the number of possible observations at the university level drops to eight. We count as direct observations backward and forward projections constructed using detailed census data on educational attainment broken down by age group and other estimates that incorporate substantial information from census or survey data.

As can be seen in Table 1, for most of the countries in the sample we have enough primary information to reconstruct reasonable attainment series covering
the whole sample period. The more problematic cases are highlighted using bold characters. In the case of Italy, the main problem is that much of the available information refers to the population over six years of age. For Denmark and Germany (at the secondary level), the earliest available direct observation refers to 1970 or later. In these two cases, we have projected attainment rates backward to 1960 using the attainment growth rates reported in OECD (1974), but we are unsure of the reliability of this extrapolation.

After estimating the breakdown of the adult population by educational level, we have calculated the average number of years of schooling, taking into account the theoretical duration of the different school cycles in each country. The results are summarized in Table 2. The table shows the position of the different countries relative to the sample average in each period, which is normalized to 100, with the countries arranged in decreasing order by school attainment in 1990. The last two rows of the table show the (unweighted) average years of schooling for the entire sample and the coefficient of variation of the same variable. Average attainment

---

**Table 1. Availability of primary data.**

<table>
<thead>
<tr>
<th></th>
<th>Secondary attainment</th>
<th>University attainment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct/ First/ Last</td>
<td>Direct/ First/ Last</td>
</tr>
<tr>
<td></td>
<td>tot. obs. observ.</td>
<td>tot. obs. observ.</td>
</tr>
</tbody>
</table>

---

5. In these calculations, we have implicitly assumed that everybody who starts a given school cycle does eventually complete it, which is clearly not the case. Hence, our figures will be biased upward and are not strictly comparable with the B&L96 or B&L00 average schooling series, which do make use of estimates of completion rates. We have not used the available data on this variable because they display the same anomalies as attainment and enrollment rates.
increases by 27.3% between 1960 and 1990 as a result of the important improvement in the educational level of the younger cohorts observed in practically all countries. The dispersion of attainment levels in the sample drops by 23.4% during the same period, with convergence in schooling becoming noticeably faster in the 1980s.

Both our series and C&S’s display considerably smoother time profiles than the different versions of the B&L data set. This is illustrated in Figure 5, where we have plotted the fitted distribution of the annualized growth rate of average years of schooling (for our OECD sample) in each data set.\textsuperscript{6} The difference in the range of this variable across data sets is enormous: Whereas our annual growth rates range between 0.09\% and 1.92\%, and those of C&S between 0.27\% and 3.27\%, B&L00’s go from −1.35\% to 6.13\%. As noted previously, we suspect that this feature may explain why these series often generate implausible results in growth regressions. A first indication of this is that the coefficient of a univariate

\textsuperscript{6} To check the sensitivity of the distribution of our data to the assumptions made to estimate attainment in countries for which there is little primary information, we have also constructed the fitted distribution for a “high quality” subsample that is obtained by removing from the original sample the nine countries that are shown in bold type in Table 1 or have 9/24 or less direct observations at either the secondary or university levels. We find that the distribution of growth rates for the high quality subsample is almost identical to the one for the full sample (see Figure 5).

---

**Table 2. Average years of schooling of the adult population (sample average = 100 in each year).**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>West Germany</td>
<td>118.5</td>
<td>120.1</td>
<td>121.6</td>
<td>121.7</td>
<td>121.7</td>
<td>122.1</td>
<td>121.7</td>
</tr>
<tr>
<td>Australia</td>
<td>117.7</td>
<td>120.6</td>
<td>122.6</td>
<td>124.0</td>
<td>125.7</td>
<td>124.2</td>
<td>121.1</td>
</tr>
<tr>
<td>Canada</td>
<td>124.1</td>
<td>123.5</td>
<td>123.2</td>
<td>123.1</td>
<td>122.9</td>
<td>121.2</td>
<td>119.7</td>
</tr>
<tr>
<td>USA</td>
<td>126.3</td>
<td>126.1</td>
<td>125.4</td>
<td>124.5</td>
<td>123.1</td>
<td>121.0</td>
<td>119.1</td>
</tr>
<tr>
<td>Switzerland</td>
<td>124.8</td>
<td>124.2</td>
<td>123.6</td>
<td>120.5</td>
<td>117.8</td>
<td>116.1</td>
<td>114.9</td>
</tr>
<tr>
<td>New Zealand</td>
<td>125.1</td>
<td>123.4</td>
<td>121.7</td>
<td>119.6</td>
<td>117.5</td>
<td>115.4</td>
<td>113.8</td>
</tr>
<tr>
<td>Denmark</td>
<td>129.0</td>
<td>125.9</td>
<td>123.0</td>
<td>119.8</td>
<td>116.9</td>
<td>113.7</td>
<td>110.2</td>
</tr>
<tr>
<td>Austria</td>
<td>107.7</td>
<td>105.4</td>
<td>103.5</td>
<td>103.2</td>
<td>104.1</td>
<td>105.9</td>
<td>106.3</td>
</tr>
<tr>
<td>Japan</td>
<td>103.1</td>
<td>103.3</td>
<td>103.5</td>
<td>104.8</td>
<td>105.8</td>
<td>105.6</td>
<td>105.6</td>
</tr>
<tr>
<td>Norway</td>
<td>115.8</td>
<td>113.6</td>
<td>111.6</td>
<td>108.9</td>
<td>107.1</td>
<td>106.1</td>
<td>104.4</td>
</tr>
<tr>
<td>Finland</td>
<td>91.5</td>
<td>94.5</td>
<td>96.8</td>
<td>98.6</td>
<td>100.7</td>
<td>102.0</td>
<td>103.1</td>
</tr>
<tr>
<td>Netherlands</td>
<td>97.0</td>
<td>97.6</td>
<td>98.1</td>
<td>99.0</td>
<td>100.1</td>
<td>101.4</td>
<td>102.9</td>
</tr>
<tr>
<td>Sweden</td>
<td>96.2</td>
<td>95.5</td>
<td>95.0</td>
<td>96.1</td>
<td>97.2</td>
<td>98.4</td>
<td>99.8</td>
</tr>
<tr>
<td>UK</td>
<td>102.5</td>
<td>101.7</td>
<td>100.8</td>
<td>99.9</td>
<td>99.0</td>
<td>98.8</td>
<td>98.9</td>
</tr>
<tr>
<td>France</td>
<td>97.3</td>
<td>98.6</td>
<td>100.2</td>
<td>101.3</td>
<td>99.9</td>
<td>98.9</td>
<td>98.2</td>
</tr>
<tr>
<td>Belgium</td>
<td>92.5</td>
<td>93.3</td>
<td>94.1</td>
<td>94.4</td>
<td>94.8</td>
<td>94.7</td>
<td>94.7</td>
</tr>
<tr>
<td>Ireland</td>
<td>88.9</td>
<td>86.8</td>
<td>86.9</td>
<td>86.5</td>
<td>86.0</td>
<td>87.0</td>
<td>88.4</td>
</tr>
<tr>
<td>Italy</td>
<td>64.7</td>
<td>66.7</td>
<td>68.6</td>
<td>69.6</td>
<td>70.7</td>
<td>73.1</td>
<td>75.6</td>
</tr>
<tr>
<td>Greece</td>
<td>66.5</td>
<td>67.5</td>
<td>68.5</td>
<td>70.1</td>
<td>71.8</td>
<td>73.1</td>
<td>74.3</td>
</tr>
<tr>
<td>Spain</td>
<td>59.5</td>
<td>58.5</td>
<td>57.5</td>
<td>58.5</td>
<td>59.5</td>
<td>62.8</td>
<td>66.7</td>
</tr>
<tr>
<td>Portugal</td>
<td>52.3</td>
<td>53.2</td>
<td>54.0</td>
<td>56.0</td>
<td>58.0</td>
<td>59.0</td>
<td>60.2</td>
</tr>
<tr>
<td>avg. (in years)</td>
<td>8.36</td>
<td>8.69</td>
<td>9.02</td>
<td>9.45</td>
<td>9.87</td>
<td>10.28</td>
<td>10.64</td>
</tr>
<tr>
<td>coeff. of variat.</td>
<td>22.8</td>
<td>22.3</td>
<td>21.8</td>
<td>20.9</td>
<td>20.1</td>
<td>18.9</td>
<td>17.5</td>
</tr>
</tbody>
</table>
regression of the growth rate of output per employed worker on the growth rate of average years of schooling (with both variables measured in deviations from their contemporaneous sample means) increases from 0.173 (with a t-ratio of 1.89) with the B&L00 data to 0.348 (with \( t = 3.07 \)) with the C&S series, and to 1.03 (with \( t = 3.16 \)) with those constructed in this paper.

4. Comparative Growth Results for Different Data Sets

In this section we explore more systematically how growth results change as we make use of the different schooling data sets that are available in the literature. We will assume that the educational attainment of employed workers (\( HE \)) is one of the inputs in a constant-returns Cobb–Douglas aggregate production function which we will write in intensive form

\[
q_{it} = a_{it} + \alpha z_{it} + b^* h e_{it},
\]

(1)

where \( q_{it} \) is the log of output per employed worker in country \( i \) at time \( t \), \( z \) the log of the stock of physical capital per employed worker, \( h e \) the log of the average number of years of schooling of employed workers and \( a_{it} \) the log of total factor productivity (TFP). One difficulty we face when trying to estimate (1) is that our attainment data (\( H \)) generally refer to the adult population rather than to employed workers. To get around this problem, we hypothesize that \( H E \) increases with population attainment and decreases with the ratio of employment to the adult population (\( E \)) so that

\[
h e_{it} = c^* h_{it} - d^* e_{it},
\]

(2)

where all variables are measured in logarithms. Substituting (2) into (1) we obtain the reduced-form production function

\[
q_{it} = a_{it} + \alpha z_{it} + \beta h_{it} - \psi e_{it},
\]

(3)
where
\[ \beta = bc \quad \text{and} \quad \varphi = bd. \quad (4) \]

Hence, our use of population data is likely to introduce a bias in the human capital coefficient that cannot be corrected without outside information on the value of the parameter \( c \). We will attempt to obtain one such estimate later on, but for now we will concentrate on the estimation of \( \beta \), keeping in mind that this coefficient is likely to be a biased estimate of \( b \). On the other hand, the reduced-form parameter \( \beta \) is probably the more relevant one from a policy point of view, because governments have a greater degree of control over the attainment level of the entire population than over that of employed workers.

We will estimate a number of specifications based on (3) using different schooling series and a common set of other variables. Our first specification is obtained directly from (3) by replacing \( a_{it} \) by a set of period dummies \( (\eta_{1t}) \),
\[ q_{it} = \Gamma_1 + \eta_{1t} + \alpha z_{it} + \beta h_{it} - \varphi e_{it} + \epsilon_{1it}, \quad (5) \]
where \( \epsilon_{1it} \) is a disturbance term. This equation implicitly assumes that all countries have access to the same technology. The second specification introduces fixed country effects \( (\gamma_i) \) to allow for cross-country differences in TFP levels,
\[ q_{it} = \Gamma_2 + \gamma_i + \eta_{2t} + \alpha z_{it} + \beta h_{it} - \varphi e_{it} + \epsilon_{2it}, \quad (6) \]
and the third one is obtained by taking (annualized) differences of (6)
\[ \Delta q_{it} = \Gamma_3 + \eta_{3t} + \alpha \Delta z_{it} + \beta \Delta h_{it} - \varphi \Delta e_{it} + \epsilon_{3it}, \quad (7) \]
where \( \Delta \) denotes annual growth rates (over the subperiod starting at time \( t \)).

Our data on output, employment, and investment in a sample of 21 OECD countries are taken from an updated version of Dabán, Doménech, and Molinas (1997), who replicate Summers and Heston’s (1991) Penn World Table for the OECD using National Accounts data from this organization and a set of purchasing power parities specific to this sample. The stock of physical capital is constructed using a perpetual inventory procedure as described in D&D02. We use pooled data at five-year intervals starting in 1960 and ending either in 1985 or in 1990 depending on the duration of the different schooling series.

The results obtained with the different specifications are reported in Table 3. The columns of the table correspond to the different attainment series that have been used in the growth literature. We work, in particular, with the estimates of

---

7. In D&D02 we also estimate a fourth specification that includes a technological catch-up effect. The results are qualitatively similar to those reported in Table 4.
8. The OECD has recently revised these series to incorporate changes in national accounting standards. As a result, our non-schooling data are slightly different from those used in de la Fuente and Doménech (2000 and 2001), and so are the results of the growth equations reported in this section.
| Panel (a): Log levels (without fixed country effects) |  
| H data from: NSD KYR B&L93 B&L96 B&L00 C&S D&D00 D&D02 avg.e. |  
| $\alpha$ | 0.580 (18.56) | 0.588 (15.77) | 0.512 (13.35) | 0.512 (14.43) | 0.479 (12.62) | 0.447 (11.81) | 0.451 (13.24) | 0.448 (12.04) | 0.502 (13.98) |  
| $\beta$ | 0.078 (2.02) | 0.186 (2.18) | 0.141 (4.49) | 0.165 (4.82) | 0.238 (6.19) | 0.397 (7.98) | 0.407 (7.76) | 0.378 (6.92) | 0.249 (5.30) |  
| $\phi$ | -0.257 (3.45) | -0.235 (3.07) | -0.311 (4.05) | -0.362 (5.19) | -0.563 (5.73) | -0.76 (7.46) | -0.923 (7.30) | -0.914 (6.97) | 0.414 (5.40) |  
| adj. $R^2$ | 0.881 | 0.884 | 0.890 | 0.879 | 0.891 | 0.918 | 0.923 | 0.914 |  
| std. error reg. | 0.129 | 0.103 | 0.124 | 0.119 | 0.113 | 0.108 | 0.105 | 0.128 |  
| no. of observ. | 126 | 95 | 126 | 147 | 147 | 147 | 147 | 147 |  

| Panel (b): Log levels with fixed country effects |  
| H data from: NSD KYR B&L93 B&L96 B&L00 C&S D&D00 D&D02 avg.e. |  
| $\alpha$ | 0.540 (16.92) | 0.533 (16.29) | 0.531 (18.59) | 0.539 (18.54) | 0.536 (19.90) | 0.516 (20.08) | 0.506 (18.17) | 0.491 (19.80) | 0.524 (18.54) |  
| $\beta$ | 0.068 (0.76) | 0.066 (1.86) | 0.136 (3.30) | 0.115 (1.80) | 0.203 (3.74) | 0.608 (4.49) | 0.627 (3.99) | 0.958 (6.51) | 0.348 (3.31) |  
| adj. $R^2$ | 0.978 | 0.980 | 0.979 | 0.977 | 0.978 | 0.980 | 0.978 | 0.982 |  
| std. error reg. | 0.056 | 0.043 | 0.054 | 0.058 | 0.057 | 0.053 | 0.056 | 0.051 |  
| no. of observ. | 126 | 95 | 126 | 147 | 147 | 147 | 147 | 147 |  
| p-value f.e. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |  

| Panel (c): Growth rates |  
| H data from: NSD KYR B&L93 B&L96 B&L00 C&S D&D00 D&D02 avg.e. |  
| $\alpha$ | 0.504 (9.78) | 0.534 (9.62) | 0.531 (9.79) | 0.539 (9.59) | 0.536 (9.71) | 0.516 (9.22) | 0.506 (9.36) | 0.491 (9.34) | 0.493 (9.21) |  
| $\beta$ | 0.079 (0.70) | 0.009 (0.15) | 0.089 (2.52) | 0.083 (1.47) | 0.079 (1.28) | 0.055 (2.03) | 0.052 (2.71) | 0.744 (7.10) | 0.266 (1.75) |  
| adj. $R^2$ | 0.978 | 0.980 | 0.979 | 0.977 | 0.978 | 0.980 | 0.978 | 0.982 |  
| std. error reg. | 0.0097 | 0.0094 | 0.0095 | 0.0096 | 0.0096 | 0.0094 | 0.0094 | 0.0093 |  
| no. of observ. | 105 | 74 | 105 | 126 | 126 | 126 | 126 | 126 |  

| Panel (d): Averages across specifications |  
| Hd data from: NSD KYR B&L93 B&L96 B&L00 C&S D&D00 D&D02 avg.e. |  
| $\beta$ | 0.075 (1.19) | 0.104 (1.53) | 0.116 (3.21) | 0.112 (2.37) | 0.155 (3.25) | 0.457 (4.21) | 0.470 (4.38) | 0.628 (4.39) |  

*Notes: All equations include period dummies. White’s heteroscedasticity-consistent $t$ ratios in parentheses below each coefficient. The average value of $t$ shown in panel d is computed respecting the sign of the $t$ ratios obtained for the different specifications; that is, for this computation we assign to each $t$ ratio the same sign as the corresponding coefficient estimate. In panel (a), $\phi$ is the coefficient of the ratio of employment to the population aged 15 to 64 ($e$).*
ratios for each data set computed across the different specifications, and the last column within each panel reports average results across data sets for each specification.

The results shown in panel (a) correspond to equation (5) and control for the employment ratio \( e_{it} \), which is highly significant and displays the expected negative sign. For the remaining equations, \( e_{it} \) turned out to be non-significant (which is not surprising given its very small time variation), so this variable is omitted with very marginal changes in the remaining coefficients. The last row of panel (b) gives the results of an \( F \) test for the exclusion of the country dummies. In all cases, the \( p \)-value is 0.00, indicating that the levels equation is almost certainly misspecified. Because country fixed effects capture permanent components of productivity that are very likely to be correlated with school attainment through a demand channel, their omission in the levels equation is likely to induce a positive (“reverse causation”) bias in the schooling coefficient. Reverse causation can be expected to be a less important problem in the other two specifications, which remove the permanent components of productivity from the disturbance term.\(^{10}\)

In the differenced specification, moreover, the period over which growth rates are computed (5 years) is short enough to make it unlikely that productivity growth will have time to feed back into the evolution of the educational stock of the adult population which, unlike enrollment rates, will respond to changes in income only with a considerable lag.

The pattern of results that emerges as we change the source of the schooling data is consistent with our hypothesis about the importance of educational data quality for growth estimates. For all the data sets, the coefficient of human capital \( (\beta) \) is positive and significant in the specification in levels without fixed country effects (panel (a) of Table 3), but the size and significance of the estimates increases appreciably as we move to the right in the table, that is, as we turn to more recent data sets where the amount of measurement error can be expected to be smaller. The differences are even sharper when the estimation is repeated with fixed country effects (panel b) or with the data in growth rates (panel c). The results obtained with the KYR, B&L, and NSD data in growth rates are generally consistent with those reported by KYR, Benhabib and Spiegel (1994), and Pritchett (1999), who find insignificant coefficients for human capital in an aggregate production function estimated with differenced data. On the other hand, our own series and those of C&S produce rather large and precise estimates of the coefficient of schooling in most equations.\(^{11}\)

\(^{10}\) Hence, we would expect estimated schooling coefficients to be higher in the levels equation than in the other two. Although this prediction does not hold in Table 4, the expected pattern will become apparent once we correct for measurement error bias taking into account the correlation of measurement error across data sets (see Section 6.2).

\(^{11}\) As a robustness check, we have reestimated all the equations reported in Table 3 using the high-quality subsample described in footnote 6. The pattern of changes in the estimated coefficient
5. Measuring Data Quality

The results reported in Table 3 show that recently constructed schooling data sets generally perform much better in growth equations than their older counterparts. Because more recent attainment series make use of previously unexploited sources of information and incorporate a number of refinements in the procedures used to estimate missing observations and/or to remove the effects of changes in classification criteria, it seems reasonable to expect that their information content will be higher than that of the older data. In this section we show that this is indeed the case using an extension of the procedure developed by Krueger and Lindhal (2001) to construct statistical indicators of data quality.

Let $H$ be the true stock of human capital and let $P_1 = H + \varepsilon_1$ be a noisy proxy for this variable, where the measurement error term $\varepsilon_1$ is an i.i.d. disturbance with zero mean and uncorrelated with $H$. The information content of this series can be measured by the ratio of signal to signal plus measurement noise in the data, that is, by the reliability ratio of the series,

$$r_1 \equiv \frac{\text{var} H}{\text{var} P_1} = \frac{\text{var} H}{\text{var} H + \text{var} \varepsilon_1}. \quad (8)$$

Assume now that in addition to $P_1$ we have a second imperfect measure of human capital, $P_2 = H + \varepsilon_2$, where $\varepsilon_2$ is also i.i.d. noise. Krueger and Lindhal (2001) show that if the error terms of the two series, $\varepsilon_1$ and $\varepsilon_2$, are not correlated with each other, then the covariance between $P_1$ and $P_2$ can be used to estimate the variance of $H$, which is the only unknown magnitude in equation (8). It follows that, under this assumption, $r_1$ can be estimated as

$$\hat{\eta} = \frac{\text{cov}(P_1, P_2)}{\text{var} P_1}, \quad (9)$$

which turns out to be the formula for the OLS estimator of the slope coefficient of a regression of $P_2$ on $P_1$. Hence, to estimate the reliability of $P_1$ we run a regression of the form $P_2 = c + r_1 P_1$. Notice, however, that if the measurement errors of the two series are positively correlated ($E\varepsilon_1\varepsilon_2 > 0$), as may be expected in many cases, $\hat{\eta}$ will overestimate the reliability ratio and hence understate the extent of the attenuation bias induced by measurement error.

of human capital across data sets is very similar to the one found in Table 3, and so are the estimates of $\beta$ obtained with the D&D02 data. The estimated values of this coefficient are 0.290 (with $t = 2.52$) for the specification in levels, 1.038 (4.23) for the levels equation with fixed effects, and 0.713 (2.14) for the differenced specification.

12. Intuitively, regressing $P_2$ on $P_1$ gives us an idea of how well $P_1$ explains the true variable $H$ because measurement error in the dependent variable ($P_2$ in this case) will be absorbed by the disturbance without generating any biases. Hence, it is almost as if we were regressing the true variable on $P_1$. 

"zwu001060311" — 2006/1/24 — page 16 — #16
In this section, we develop two extensions of this procedure that exploit the availability of multiple schooling series to improve the efficiency of Krueger and Lindhal’s “pairwise” estimator of the reliability ratio and to relax some of these authors’ rather restrictive assumptions about the nature of the measurement error term.

We will construct two alternative estimators for reliability ratios. As in Krueger and Lindhal, the first estimator assumes that measurement error is uncorrelated across data sets and is constructed by using a given data set $P_j$ to try to explain alternative estimates of the same variable ($P_k$ with $k \neq j$). The main difference is that, rather than running a single pairwise regression with just two data sets, we can increase the efficiency of the procedure by using several different schooling series as dependent variables in a restricted SUR model of the form

$$P_k = c_k + r_j P_j + u_k \quad \text{for} \quad k = 1, \ldots, K \text{ with } k \neq j,$$

(10)

where $k$ denotes the “reference” data set and varies over the last available version of all data sets different from $j$. The reliability ratio of B&L00’s data set, for instance, is estimated by using these authors’ estimate of average years of schooling as the explanatory variable in a restricted set of regressions where the reference (dependent) variables are the average years of schooling estimated by KYR, NSD, C&S, and ourselves. Other versions of the B&L data set, however, are not used as a reference because the correlation of measurement errors across the same family of schooling series is almost certainly very high and this will artificially inflate the estimated reliability ratio. We will denote the estimate of $r_j$ obtained in this way by $\hat{r}_j$, and refer to it as the SUR reliability ratio.

5.1. Estimating Reliability Ratios with Correlated Errors

Our second estimator will allow measurement error to be correlated across data sets. Given a set of noisy proxies for the stock of human capital,

$$P_j = H + \varepsilon_j \quad \text{with} \quad j = 1, 2, \ldots, J,$$

(11)

we will assume that the measurement error terms have the following structure:13

$$\varepsilon_j = \omega_j + \rho_j \varepsilon,$$

(12)

where $\omega_j$ is an idiosyncratic error component and $\rho_j$ a coefficient that measures the extent to which data set $j$ amplifies or dampens a common source of error (say, that present in the underlying primary data) which is captured by an

13. In D&D02 we consider a more general setting where we allow measurement error to be potentially correlated with all the regressors of the growth equation except for $H$ itself. We assume, in particular, that $\varepsilon_j = \omega_j + \rho_j \varepsilon + X\delta_j$, where $X = (Z, E)$ and $\delta_j = (\delta_{jz}, \delta_{je})'$. This extension introduces a number of technical difficulties that considerably complicate the exposition without qualitatively changing the results.
i.i.d. disturbance, $\varepsilon$. Notice that this specification allows measurement error to be correlated across data sets (as $E_{j,k}\varepsilon_j\varepsilon_k$ will be different from zero if $\rho_{j,k} \neq 0$). We will assume that both the common and the idiosyncratic components of the measurement error terms are uncorrelated with each other and with $H$, that is,

$$EH\varepsilon = E_{Hj}\omega_j = E_{o_j}\varepsilon = E_{o_j}\omega_k = 0 \quad \text{for all } j \text{ and } k \neq j.$$  (13)

We will also assume that the components of the measurement error term are uncorrelated with the dependent variable of the growth equation, $Q$, and with the vector $X = (Z, E)$ of non-schooling regressors, so that

$$EQ\varepsilon = EQ_{o_j} = EX_n\varepsilon = EX_n\omega_j = 0 \quad \text{for all } j \text{ and for all components } X_n \text{ of } X.$$  (14)

Under these assumptions, the procedure developed by Krueger and Lindhal (2001) will generally yield asymptotically biased estimates of reliability ratios. Consistent estimates of these ratios, however, can be recovered through a two-step procedure. We will first obtain all possible pairwise estimates of the reliability ratios of the different schooling data sets by regressing pairs of these series on each other as suggested by Krueger and Lindhal. As part of the first stage, we will also estimate a set of auxiliary regressions that will exploit the information contained in the correlations between the different schooling series and the remaining regressors of the growth equation. The estimates obtained in this manner will be biased, but because the bias will depend in a systematic way on the parameters of the measurement error process, they can be used as data in a set of second-stage regressions that will yield consistent estimates of the error parameters and of the reliability ratios of the different data sets. These second-stage equations will be derived from the probability limits of the first-stage OLS estimators.

As noted, the first step in the estimation procedure involves regressing the different schooling series on each other and on the components of $X = (Z, E)$, that is, on the stock of physical capital per worker and on the employment ratio. We fix some data set $P_j$ and use it to try to explain the remaining data sets $k \neq j$, as well as the other variables contained in the vector $X$. Hence, for each $j$ we estimate by OLS the following set of equations:

$$P_k = c_{jk} + r_{jk}P_j + u_{jk} \quad \text{for } k = 1, \ldots, J \text{ with } k \neq j \quad \text{and}$$  (15)

$$X_n = c_{jn} + \mu_{jn}P_j + u_{jn} \quad \text{for } n = z, e,$$  (16)

where the $u$'s are disturbance terms, the $c$'s are constants, and $J (= 8)$ the number of alternative proxies for $H$ that are available. 14 This yields (inconsistent) esti-

---

14. As elsewhere in this paper, these equations include period dummies and, where appropriate, country dummies as well in order to remove time and/or country means.
mates of $r_j$ and of the parameters $\mu_n$ that would be obtained by regressing the components of $X$ on the correctly measured stock of human capital, $H$. We will denote these first-round estimates by $\hat{r}_{jk}$ and $\hat{\mu}_{jn}$. (Hats will be used throughout to indicate possibly biased first-stage OLS estimates and tildes will be reserved for second-stage estimates and for other consistent estimates of various quantities.)

To derive the second-stage regressions, we begin by calculating the probability limits of the first-stage estimators as the number of observations in each schooling data set goes to infinity. It is shown in the Appendix that

$$\text{plim} \hat{r}_{jk} = r_j (1 + e_j e_k)$$

and

$$\text{plim} \hat{\mu}_{jn} = r_j \mu_n,$$

where $\mu = (\mu_z, \mu_e)'$ and $r_j$ are the true values of the parameters of interest and $e_j$ is given by

$$e_j \equiv \rho_j \sqrt{\frac{E\varepsilon^2}{EH^2}}.$$  (19)

Hence, the pairwise estimate of the reliability ratio of series $P_j$ obtained using $P_k$ as a reference, $\hat{r}_{jk}$ will be biased upward (downward) whenever the measurement error components of the two series are positively (negatively) correlated.

Equations (17) and (18) relate the expected values of $\hat{r}_{jk}$ and $\hat{\mu}_{jn}$ in large samples to the true values of $r_j$ and $\mu_n$ and to the coefficients that describe the structure of the measurement error terms. We construct the second-stage regressions as the natural finite-sample approximations to these asymptotic relations. That is, we suppress the probability limits in the left-hand side of (17) and (18), leaving as dependent variables the actual first-stage parameter estimates, and introduce a disturbance term to capture small-sample deviations from the asymptotic relationship between OLS estimates and the relevant parameters.

Thus, the equations to be estimated in the second stage can be written as

$$\hat{r}_{jk} = r_j (1 + e_j e_k) + \eta_{jk}$$

and

$$\hat{\mu}_{jn} = r_j \mu_n + \eta_{jn},$$

where $\eta_{jk}$ and $\eta_{jn}$ are disturbance terms. We estimate both equations jointly by “stacking them” so that, for each $j$, the first $J$ observations of the dependent variable (one of which will be missing as $k$ must be different from $j$) correspond to the pairwise estimates of the reliability ratio, $\hat{r}_{jk}$, and the last two observations correspond to the first-stage estimates, $\hat{\mu}_{jn}$.\(^\text{15}\) We will denote by $\tilde{r}_j$ the reliability

\(^{15}\) Notice that the system formed by equations (18) and (19) is non-linear. For the Eviews NLS algorithm to start iterating, non-zero initial values must be assigned to at least some of the parameters.
5.2. Results for the Different Data Sets

We have constructed SUR and consistent estimates of reliability ratios for the different transformations of the schooling data sets used in Section 4. Restricting ourselves to the sample of 21 OECD countries covered by our own series, we have estimated reliability ratios for the data in logs \((h_{it})\) and in annual growth rates \((\Delta h_{it})\), and for the “within” transformation used to remove fixed effects \((h_{it} - h_i)\). Table 4 summarizes the results. Estimated reliability ratios are shown in relative terms, with the average value for each column normalized to 100. The three blocks of the table refer to the different growth specifications we have used (levels, fixed effects, and differences or growth rates). At the bottom of each block of columns we show the average value (across data sets) of the reliability ratios and the correlation between the two alternative estimates (SUR and consistent) of this variable.

Inspection of the table reveals a number of interesting results. The first one is that the information content of the schooling data is quite low on average. When fixed effects are removed or we work with differenced data, our mean estimate of the reliability ratio (for all series and both transformations) lies between 0.31 and 0.37, depending on whether we use the SUR or the consistent estimates. Things appear to be better for the data in levels, but only if we restrict ourselves to the SUR estimates (whose average value is 0.617). Once we turn to the consistent estimates, the mean reliability ratio drops by two-thirds to 0.221, which suggests that SUR reliability ratios are highly inflated by the positive correlation of ratio estimates obtained from the second-stage equations (20) and (21) and we will refer to them as consistent reliability ratios (because they will be consistent estimators of \(r_j\) in the presence of correlated errors).

We set initial values by estimating a log-linear approximation to the first equation. We have also repeated the estimation in RATS and obtained very similar results. (See the Appendix to D&D02 for additional details on the estimation procedure.)

---

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Levels</th>
<th>Fixed effects</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUR consistent</td>
<td>SUR consistent</td>
<td>SUR consistent</td>
</tr>
<tr>
<td>NSD</td>
<td>53.6</td>
<td>25.9</td>
<td>26.0</td>
</tr>
<tr>
<td>KYR</td>
<td>97.3</td>
<td>30.2</td>
<td>25.5</td>
</tr>
<tr>
<td>B&amp;L93</td>
<td>70.7</td>
<td>33.1</td>
<td>11.5</td>
</tr>
<tr>
<td>B&amp;L96</td>
<td>83.6</td>
<td>39.9</td>
<td>15.5</td>
</tr>
<tr>
<td>B&amp;L00</td>
<td>97.7</td>
<td>48.5</td>
<td>17.6</td>
</tr>
<tr>
<td>C&amp;S</td>
<td>147.9</td>
<td>149.2</td>
<td>184.1</td>
</tr>
<tr>
<td>D&amp;D00</td>
<td>123.4</td>
<td>223.0</td>
<td>216.8</td>
</tr>
<tr>
<td>D&amp;D02</td>
<td>125.8</td>
<td>250.2</td>
<td>303.0</td>
</tr>
<tr>
<td>avg. value</td>
<td>0.617</td>
<td>0.367</td>
<td>0.254</td>
</tr>
<tr>
<td>corr. w/SUR</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

SUR estimates assume that measurement error is uncorrelated across data sets; consistent estimates relax this assumption.
measurement error across the different series in levels. Because reliability ratios must lie between zero and one (with zero indicating that the series contains no information and one that it is measured without error) these figures imply that, for a sample of OECD countries, more than two-thirds of the variation in the schooling data used to estimate the average growth equation is noise.

A more encouraging finding is that recent efforts to increase the signal content of the schooling data seem to have been rather successful. Taking as a reference the average value of the consistent reliability ratio across data transformations, the information content of the three data sets listed at the bottom of Table 4 is between three and four times larger than that of the B&L96 schooling series.

On the whole, our estimates of reliability ratios display a fairly consistent picture of the relative quality of the different series. There are, however, some patterns in the results that suggest that our assumptions about the nature of measurement error may not be adequate for the estimation of the reliability ratios of some data sets. The clearest example is the behaviour of the relative reliability ratios for the KYR and NSD series in levels, which rise sharply as we go from the SUR to the consistent estimates. We suspect that this anomaly may be due to our assumption of a common source of error arising from reliance on similar primary data. Because the KYR and NSD series rely more heavily on enrollment data than the rest of the data sets, their estimated correlation with the primary error shared by the remaining series (captured by the parameter $\rho_j$) is likely to be quite low regardless of the amount of error contained in the alternative primary sources used in these papers. This, in turn, will cause the consistent reliability ratios for these two series to increase relative to those of the remaining data sets.

6. Correcting for Measurement Error Bias

The results of the previous two sections show that there is a systematic relationship between the quality of schooling data and OLS estimates of the coefficient of human capital in an aggregate production function. To make this point visually, Figure 6 plots the various estimates of $\beta$ given in panels (a) through (c) of Table 3 against the relevant SUR reliability ratios, along with the regression lines fitted for each of the growth specifications. The scatter shows a strong positive correlation between these two variables within each specification. It also suggests that the observed relationship between reliability ratios and OLS coefficients may be extrapolated to the limiting case of no measurement error in order to obtain consistent estimates of the schooling parameter—and that such an exercise will lead to the conclusion that the true value of $\beta$ is at least 0.50 (which is the prediction of the levels equation for $r = 1$).

In the remainder of this section we will use an extension of the classical errors-in-variables model to show that the extrapolation procedure suggested by Figure 6 is essentially correct, provided we adjust reliability ratios to reflect the further loss
of signal that takes place when schooling is not the only regressor in the growth equation. We will then use this approach to construct a set of meta-estimates of \( \beta \) that will be free of the attenuation bias induced by measurement error.

We will write the model we want to estimate (that is, the different versions of the growth equation discussed in Section 4) in the generic form

\[
Q = H\beta + X\alpha + u_1,
\]

where \( Q \) is (some transformation of) output per employed worker, \( H \) the true stock of human capital, \( X = (Z, E) \) a row vector whose components are the other growth regressors (the stock of capital per employed worker, \( Z \), and the employment ratio, \( E \)) and \( \alpha \) a column vector of coefficients. It will be assumed that the error term \( u_1 \) satisfies all the standard assumptions of the linear regression model (and, in particular, that it is uncorrelated with the regressors) so that the estimation of (22) by OLS with the correctly measured stock of human capital would be consistent.

We want to calculate the bias that arises when \( \beta \) is estimated in equation (22) using a noisy proxy for the stock of human capital, \( P_j = H + \varepsilon_j \), rather than \( H \) itself. Under the assumptions about the nature of the measurement error term given in equations (12)–(14) in Section 5, it is shown in the Appendix that the probability limit of the OLS estimator of \( \beta \) obtained using schooling series \( P_j \) (calculated as the number of observations in the series goes to infinity) is given by\(^{16}\)

\[
\text{plim} \hat{\beta}_j = \frac{r_j - ER_j^2}{1 - ER_j^2} \beta (23')
\]

\(^{16}\) Equation (23) can be shown to be equivalent to the bias-correction formula given in Krueger and Lindahl (2001), which, in our notation, would be written

\[
\text{plim} \hat{\beta}_j = \frac{r_j - ER_j^2}{1 - ER_j^2} \beta (23')
\]
where $\beta$ is the true value of the schooling parameter, $ER^2_H \equiv \text{plim} \ R^2(H|X)$ is the probability limit of the coefficient of determination of a regression of $H$ on the rest of the explanatory variables of the growth equation, $X$, and $r_j$ is the reliability ratio of series $P_j$.

We will refer to the expression that multiplies the true value of the parameter in the right-hand side of (23) as the attenuation coefficient of series $j$, and we will denote it by $a_j$. Notice that the attenuation coefficient reduces to the reliability ratio when $P_j$ is the only regressor in the growth equation (or when $H$ is perfectly uncorrelated with the rest of the right-hand side variables). Otherwise, there is a further loss of signal that increases with the correlation of $H$ with the other explanatory variables in the model. Using the fact that $ER^2_H$ and $r_j$ must both lie between zero and one, it is easily seen that $a_j$ is an increasing function of $r_j$ with $0 \leq a_j \leq r_j \leq 1$, and that $a_j < r_j$ except if $r_j = 0$, $r_j = 1$ or $ER^2_H = 0$. Hence, measurement error induces a bias towards zero, but it cannot reverse coefficient signs in large samples, even when the human capital proxy is highly correlated with other regressors. This last result is not immediately apparent in the alternative (and equivalent) bias correction formula given in Krueger and Lindhal (2001) (equation (23′) in footnote 16).

Equation (23) generalizes the well-known result that the attenuation bias induced by classical measurement error (which would correspond to the special case where $\rho_j = 0$ for all $j$) is inversely proportional to a generalized measure of the information content of the variable of interest that takes into account its correlation with other regressors. This expression implies that the extrapolation procedure proposed in the previous section will work correctly, provided we replace the reliability ratio with the attenuation coefficient. It also implies that, if the attenuation coefficients of the different series are correctly estimated, the regression line that describes the relationship between this variable and the corresponding OLS estimates of $\beta$ will go through the origin. This suggests the following strategy for constructing meta-estimates of $\beta$. First, we will estimate (for each of the different specifications of the production function we have used) a regression of the form

$$\hat{\beta}_j = \kappa + \beta \tilde{a}_j + \eta_j,$$

(24)

where $\hat{\beta}_j$ is the OLS estimate of $\beta$ obtained using schooling series $P_j$, $\tilde{a}_j$ a consistent estimate of the relevant attenuation coefficient, $\kappa$ a constant, and $\eta_j$ a disturbance term. If the hypothesis that $\kappa = 0$ cannot be rejected, we will drop the constant and obtain our meta-estimate of $\beta$ as the slope coefficient of the restricted regression. If the hypothesis can be rejected, we must conclude that there is something wrong with our estimates of the attenuation coefficients, for
equation (24) would then imply that regressing $Q$ on pure noise will produce a non-zero coefficient.

Once we drop the constant, equation (24) is the finite-sample counterpart of equation (23). It is obtained from this last expression by deleting the probability limit, leaving on the left-hand side the corresponding sample estimate, and by replacing the attenuation coefficient on the right-hand side with a consistent estimate of it and adding a disturbance term.

6.1. Estimation of Attenuation Coefficients

The values of the dependent variable for equation (24) are given in Table 3. The values of the regressor, however, have to be constructed using the results given in Table 4 and a consistent estimate of $ER_{H}^2$. It is shown in the Appendix that such an estimate can be constructed as

$$ER_{H}^2 = \hat{\mu}^\prime \hat{\phi},$$

(25)

where $\hat{\mu}$ and $\hat{\phi}$ are consistent estimates of the vectors of coefficients that would be obtained by regressing the components of $X$ on the correctly measured stock of human capital, $H$, and from the “reverse” regression of $H$ on $X$. The first of these vectors has been estimated in Section 5.1 and an efficient estimate of the second one can be obtained by running a system of the form

$$P_j = X\phi + u_{xj} \quad \text{for } j = 1, \ldots, J,$$

(26)

where we restrict the vector of coefficients $\phi$ to be equal for all equations. An alternative estimate of $ER_{H}^2$ can be obtained using the following relation:

$$ER_{j}^2 = r_j ER_{H}^2,$$

(27)

which is also derived in the Appendix.

Table 5 reports our estimates of attenuation coefficients ($a_j$) under the same assumptions about the nature of the measurement error process that underlie the SUR and consistent reliability ratios given in Table 4. The format is the same as in the previous table: We report relative values for each data set and, at the bottom of the table, the average value of the attenuation coefficient and its correlation with the SUR reliability ratio. The last two rows of the table contain, respectively, the value of $\hat{\mu}^\prime \hat{\phi}$ and the value of $ER_{H}^2$ that was used to construct the displayed attenuation coefficients.

The value of $a_j$ is calculated using the definition of this variable given in equation (23) and an estimate of $ER_{H}^2$. For the case of the consistent reliability ratios, $ER_{H}^2$ is generally obtained as $\hat{\mu}^\prime \hat{\phi}$ using equation (25) except in one case when this procedure yields an estimate of $ER_{H}^2$ that does not lie between zero and one (see the note to Table 5). In the case of the SUR reliability ratios, the value of $ER_{H}^2$ cannot be obtained in the same way because no estimate $\mu$ is then available.
Table 5. SUR and consistent attenuation coefficients for different data sets, relative values.

<table>
<thead>
<tr>
<th></th>
<th>Levels</th>
<th></th>
<th>Fixed effects</th>
<th></th>
<th>Differences</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUR</td>
<td>consistent</td>
<td>SUR</td>
<td>consistent</td>
<td>SUR</td>
<td>consistent</td>
</tr>
<tr>
<td>NSD</td>
<td>38.6</td>
<td>80.4</td>
<td>22.9</td>
<td>25.4</td>
<td>25.2</td>
<td>25.6</td>
</tr>
<tr>
<td>KYR</td>
<td>88.1</td>
<td>186.0</td>
<td>26.8</td>
<td>43.2</td>
<td>24.8</td>
<td>16.7</td>
</tr>
<tr>
<td>B&amp;L93</td>
<td>55.4</td>
<td>48.2</td>
<td>29.4</td>
<td>20.5</td>
<td>11.2</td>
<td>16.8</td>
</tr>
<tr>
<td>B&amp;L96</td>
<td>70.1</td>
<td>60.1</td>
<td>35.7</td>
<td>48.8</td>
<td>15.0</td>
<td>44.6</td>
</tr>
<tr>
<td>B&amp;L00</td>
<td>88.7</td>
<td>66.7</td>
<td>43.6</td>
<td>53.0</td>
<td>17.1</td>
<td>52.8</td>
</tr>
<tr>
<td>C&amp;S</td>
<td>190.5</td>
<td>123.9</td>
<td>145.3</td>
<td>139.0</td>
<td>183.0</td>
<td>221.7</td>
</tr>
<tr>
<td>D&amp;D00</td>
<td>132.0</td>
<td>119.5</td>
<td>231.0</td>
<td>228.4</td>
<td>216.7</td>
<td>172.2</td>
</tr>
<tr>
<td>D&amp;D02</td>
<td>136.8</td>
<td>115.2</td>
<td>265.3</td>
<td>241.7</td>
<td>307.0</td>
<td>249.7</td>
</tr>
<tr>
<td>avg. value</td>
<td>0.421</td>
<td>0.116</td>
<td>0.339</td>
<td>0.393</td>
<td>0.247</td>
<td>0.325</td>
</tr>
<tr>
<td>corr. w/SUR</td>
<td>0.984</td>
<td>0.498</td>
<td>0.999</td>
<td>0.995</td>
<td>1.000</td>
<td>0.962</td>
</tr>
<tr>
<td>reliability</td>
<td>(\hat{\mu} \hat{\phi})</td>
<td>1.069</td>
<td>0.163</td>
<td>0.039</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(ER_H^2)</td>
<td>0.607</td>
<td>0.551</td>
<td>0.198 ; 0.163</td>
<td>0.060 ; 0.039</td>
<td></td>
</tr>
</tbody>
</table>

Note: Because the value of \(\hat{\mu} \hat{\phi}\) exceeds 1 in the consistent case for the data in levels, we use as \(ER_H^2\) the highest observed value of \(R^2\). Under the assumptions used in this case, equation (27) holds so we should have \(ER_H^2 < ER_H^2\) for any \(P_j\) with \(r_j < 1\) and this implies that the reported \(\hat{\alpha}_j\)'s should overestimate the true ones. We do not use equation (27) to estimate \(ER_H^2\) because this procedure also yields a value of this parameter greater than 1.

We then rely on equation (27) and obtain an estimate of \(ER_H^2\) by regressing the observed value of \(R^2_j = R^2_j(P_j|X)\) on the SUR estimate of \(r_j\) without a constant. The average attenuation coefficient for all series and data transformations is 0.336 according to our SUR estimates and 0.274 when we correct for correlated errors. These figures become 0.146 and 0.118 when we drop the three data sets listed at the bottom of Table 5 (C&S, D&D00, and D&D02). This implies that the average estimate of the coefficient of schooling in a growth equation is likely to suffer from a very substantial downward bias and strengthens Krueger and Lindhal’s (2001) conclusion that the information content of the schooling series available a few years ago is low enough to explain some of the most widely cited negative results on human capital and growth as the result of attenuation bias.

Comparing Tables 4 and 5 and looking at the correlations given in the lower part of the second table we see that, except in the case when the data are used in levels, the corrections required to adjust for correlated errors and for the presence of additional growth regressors are very small. Hence, in fixed effects or differenced specifications, the SUR reliability ratios (which are much easier to compute than the rest of our quality indicators) are rather good measures of the information content of the different series and can be used to correct in a simple way for measurement error bias.

6.2. Meta-estimates of \(\beta\)

Table 6 shows the meta-estimates of \(\beta\) obtained through the estimation of equation

\[
\hat{\beta}_j = \kappa + \beta \hat{\alpha}_j + \eta_j,
\]
Table 6. Meta-estimates of $\beta$.

Panel (a): Assuming uncorrelated measurement error
(with SUR attenuation coefficients)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>levels</td>
<td>0.569</td>
<td>0.966</td>
<td>0.965</td>
<td>0.587</td>
<td>0.998</td>
<td>1.016</td>
</tr>
<tr>
<td>(6.30)</td>
<td>(10.05)</td>
<td>(16.31)</td>
<td>(16.50)</td>
<td>(16.12)</td>
<td>(22.23)</td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>0.009</td>
<td>0.020</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.22)</td>
<td>(0.46)</td>
<td>(1.28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.869</td>
<td>0.944</td>
<td>0.978</td>
<td>0.868</td>
<td>0.942</td>
<td>0.972</td>
</tr>
</tbody>
</table>

Panel (b): Allowing for correlated measurement error
(with consistent attenuation coefficients)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>levels</td>
<td>0.954</td>
<td>3.224</td>
<td>0.898</td>
<td>0.872</td>
<td>1.967</td>
<td>2.606</td>
<td>0.889</td>
<td>0.843</td>
</tr>
<tr>
<td>(1.03)</td>
<td>(3.77)</td>
<td>(7.76)</td>
<td>(12.52)</td>
<td>(5.31)</td>
<td>(10.05)</td>
<td>(12.80)</td>
<td>(18.82)</td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>0.138</td>
<td>-0.070</td>
<td>-0.001</td>
<td>-0.018</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1.19)</td>
<td>(0.76)</td>
<td>(0.09)</td>
<td>(0.58)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dummy Kyr</td>
<td>-0.438</td>
<td></td>
<td></td>
<td></td>
<td>-0.375</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3.46)</td>
<td>(0.09)</td>
<td></td>
<td></td>
<td>(4.07)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.151</td>
<td>0.749</td>
<td>0.909</td>
<td>0.963</td>
<td>-0.050</td>
<td>0.721</td>
<td>0.909</td>
<td>0.961</td>
</tr>
</tbody>
</table>

Note: $t$ ratios in parentheses below each coefficient.

which is reproduced here for convenience. The table has two panels that correspond to the SUR and consistent attenuation coefficients. Within each panel, the columns correspond to the three growth specifications we have been working with (levels, fixed effects, and differences). In panel (b), we add an additional specification in levels that includes a dummy for the KYR data set, which is a rather atypical observation in this case. This fact is illustrated in Figure 7, where we show the scatter diagram obtained by plotting the OLS estimates of $\beta$ against the corresponding consistent attenuation coefficients, identifying the levels observation corresponding to the KYR data set.

We first estimate equation (24) with a constant, which is generally not significant as predicted by the model and then without the constant to obtain our final estimates of the coefficient of schooling. These meta-estimates of $\beta$ consistently display large values that significantly exceed those found in the previous literature. Restricting ourselves to the results obtained with the fixed effects and the differenced growth equations, which are less likely to be misspecified and do not display obvious outliers, our estimates of $\beta$ range from 0.843 to 1.016.

Comparing panels (a) and (b) of Table 6, we see that allowing for correlated measurement error makes little difference for the meta-estimates of $\beta$ that are based on the results of fixed effects or differenced specifications. Things are very different, however, for the levels specification, where allowing for correlated errors more than triples our final estimate of $\beta$ and turns the levels coefficient from the smallest into the largest of all meta-estimates. Notice that after this
correction the ranking of the meta-estimates is consistent with our priors about the relative importance of reverse causation bias in the three specifications.

Before commenting further on these results, it should be recalled that our use of data on population attainment (rather than on the attainment of employed workers) potentially introduces a bias. As noted in Section 4, the production function parameter we would like to estimate is not $\beta$ itself, but rather $b = \beta/c$, where $c$ is one of the coefficients of a regression of the form

$$h_{eit} = c^* h_{it} - d^* e_{it},$$

(29)

where $h$ is the log of population attainment, $he$ the log of the average attainment of employed workers, and $e$ the log of the ratio of employment to the adult population. Because the data sets we have used do not provide information on the schooling of employed workers, we have used panel data for the Spanish regions from Mas et al. (2002) to estimate equation (2) and obtain an outside estimate of $c$ that may give us an idea of the size of the required correction. Because measurement error is almost certainly a problem in these data as well (and there are no alternative sources that can be used to estimate reliability ratios), we obtain a set of bounds on the value of $c$ by estimating both (2) and the reverse regression with the roles of $h$ and $he$ interchanged. The exercise is repeated with the data in (log) levels with fixed effects and in growth rates to obtain two alternative intervals of plausible values of $c$ that can be matched with the more reliable of our growth specifications to obtain ranges of possible values of $b$.

The results are shown in Table 7. The first two rows of the table show the bounds on the value of $c$ obtained from the direct and reverse regressions based on (2). The third row shows the estimates of $\beta$ taken from Table 6, and the last

![Figure 7. Estimated $\beta$ vs. consistent attenuation coefficient.](image-url)
Table 7. Range of meta-estimates of $b$.

<table>
<thead>
<tr>
<th></th>
<th>SUR</th>
<th>Consistent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fixed effs.</td>
<td>diffs.</td>
</tr>
<tr>
<td>$\hat{c}$, direct</td>
<td>0.867</td>
<td>0.983</td>
</tr>
<tr>
<td>$\hat{c}$, reverse</td>
<td>1.182</td>
<td>1.468</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>0.998</td>
<td>1.016</td>
</tr>
<tr>
<td>$b_{\min}$</td>
<td>0.844</td>
<td>0.692</td>
</tr>
<tr>
<td>$b_{\max}$</td>
<td>1.151</td>
<td>1.034</td>
</tr>
</tbody>
</table>

Two rows display the minimum and maximum values of $b$, which are obtained by dividing the estimated $\hat{\beta}$ by each of the estimates of $c$. As can be seen in the table, the correction is potentially important in some cases but does not qualitatively change our previous results. Our smallest lower bound for $b$ (0.574) is almost twice as large as Mankiw, Romer, and Weil’s (1992) estimate of $1/3$, which could probably have been considered a consensus value for this parameter a few years ago and has lately come to be seen as too optimistic in the light of recent negative results in the literature.

We interpret our results as a clear indication that investment in human capital is an important growth factor whose effect on productivity has been underestimated in previous studies because of poor data quality. Considerable uncertainty remains, however, regarding the precise value of the coefficient of years of schooling in an aggregate production function. Values of $b$ of the order of 0.6 imply Mincerian rates of return to schooling at the individual level (of between 8% and 9%) that are consistent with the evidence available from microeconometric wage equations.\(^{17}\) Somewhat higher values of this parameter remain plausible because our macroeconometric estimates should pick up any externalities that may be

\(^{17}\) As discussed in de la Fuente (2003), macroeconometric estimates of schooling parameters have to be adjusted before they can be compared with Mincerian estimates based on individual-level data (essentially because the latter do not hold the stock of capital constant). In a competitive economy where output is given by a constant-returns Cobb–Douglas function of physical and (total) human capital and where a worker’s stock of human capital is related to average schooling through a Mincerian (exponential) function, the relationship between the schooling coefficient in the aggregate production function ($b$) and the individual-level Mincerian returns parameter ($\theta$) is given by $\theta = (b/H)/(1 - \alpha)$, where $H$ is average attainment in years and the elasticity of aggregate output with respect to the stock of physical capital.

Assuming $\alpha = 1/3$, the smallest of the lower bounds for $b$ given in Table 8 (0.574) implies a value of $\theta = 8.05\%$ for the average country in our sample and of $8.7\%$ for the average EU country. These figures lie between Harmon, Walker, and Westergaard-Nielsen’s (2001) mean estimate of 8.03% for the same magnitude in a sample of EU countries and their average estimate of 10.6% for the UK and Ireland, where the estimated value of the Mincerian parameter presumably captures the contribution of education to productivity more accurately than in continental Europe due to the greater flexibility of these countries’ labour markets.

Taken as a whole, our meta-estimates of the schooling coefficient suggest that educational externalities may be important at the country level. This stands in contrast with (but does not necessarily contradict) the findings of a set of papers (see for instance Acemoglu and Angrist (2000) and Ciccone and Peri (2004)) that find no convincing evidence of such externalities at the level of cities or regions once the likely endogeneity bias induced by free labour mobility is taken into account. We interpret these results as an indication that the external benefits of education are likely to operate
associated with the accumulation of human capital, such as those arising from
the contribution of an educated labour force to the development and adoption of
new technologies. Some of our levels estimates seem decidedly implausible but,
as we have already argued, these results are likely to contain a strong upward bias
arising from reverse causation and should probably be disregarded.

7. Conclusion

Existing data on educational attainment contain a considerable amount of noise.
Because of changes in classification criteria and other inconsistencies in the
primary data, the most widely used schooling series often display implausible
time-series and cross-section profiles. After documenting some of these anoma-
lies, we have constructed new estimates of educational attainment for a sample
of OECD countries. We have attempted to increase the signal-to-noise ratio in
these data by using previously unexploited information and by eliminating sharp
breaks in the series that must reflect changes in data collection criteria.

We have also constructed statistical measures of the information content of
eight different schooling data sets under alternative assumptions on the nature of
the measurement error they contain. Although the choice of assumptions does
make some difference for the ranking of the different data sets, on the whole,
these indices support our hypothesis that the amount of measurement error in
these data is rather large and clearly suggest that both our attainment series and
those constructed by Cohen and Soto (2001) constitute a significant improvement
over earlier sources.

This paper was originally motivated by the view that weak data are likely
to be one of the main reasons for the discouraging results obtained in the recent
empirical literature on human capital and growth. Our results clearly support this
view, as does recent work by Krueger and Lindhal (2001) and Cohen and Soto
(2001), and suggest that the contribution of investment in education to productivity
growth is sizable. Our analysis of the performance of different schooling data sets
in a variety of production function specifications shows a clear tendency for human
capital coefficients to rise and become more precise as the information content of
the schooling data increases. We have extrapolated this tendency to estimate the
value of this parameter that would be obtained with the correctly measured stock
of human capital. The exercise suggests that the true value of the elasticity of
output with respect to years of schooling is almost certainly above 0.60, that is, at
least twice as large as the largest estimate of reference in the previous literature.
Appendix

In this appendix we develop an extension of the classical errors-in-variables model that will be used to construct refined estimates of reliability ratios for the different schooling series and to obtain meta-estimates of $\beta$ corrected for attenuation bias. To simplify the notation, we will assume that the distributions of the variables of interest are known, so we can work directly with population moments. The results obtained in this manner will then apply to finite samples as probability limits. It will also be assumed throughout that all the variables have zero means, so that regression constants vanish. This assumption involves no loss of generality and is, in any event, satisfied in our case, because the inclusion of time dummies in all our growth specifications is equivalent to removing period means.

We will write the model we want to estimate (the different versions of the growth equation) in the generic form

$$Q = H\beta + X\alpha + u_1, \quad (A.1)$$

where $H$ is the true stock of human capital, $X = (X_1, X_2, \ldots, X_N)$ a row vector of other regressors, and $\alpha$ a column vector of coefficients.

It will be assumed that the error term $u_1$ satisfies all the standard assumptions of the linear regression model so that the estimation of (A.1) by OLS with the correctly measured stock of human capital will be consistent. Hence, the probability limit of the OLS estimator of $\beta$ will be equal to the true value of the coefficient when $H$ is correctly measured, that is,

$$\text{plim} \hat{\beta}_H = \frac{EH'Q - EH'X(EX'X)^{-1}EX'Q}{EH'H - EH'X(EX'X)^{-1}EX'H} = \beta. \quad (A.2)$$

In practice, of course, we do not observe $H$ but only a set of noisy proxies for it,

$$P_j = H + \varepsilon_j, \quad (A.3)$$

with $j = 1, \ldots, J$, where $\varepsilon_j$ is a measurement error term. We want to calculate the bias in $\beta$ that arises when equation (A.1) is estimated using $P_j$ instead of $H$ and to estimate the reliability ratio of $P_j$, which is defined as

$$r_j \equiv \frac{EH^2}{EP_j^2}. \quad (A.4)$$

We will assume that the measurement error terms, $\varepsilon_j$, have the following structure:

$$\varepsilon_j = \omega_j + \rho_j \varepsilon, \quad (A.5)$$

where $\omega_j$ is an idiosyncratic error component and $\rho_j$ a coefficient that measures the extent to which data set $j$ amplifies or dampens a common source of error.
which is captured by an i.i.d. disturbance, \( \varepsilon \). Finally, it will be assumed that both the common and the idiosyncratic components of measurement error are uncorrelated with each other and with \( H, Q \) and \( X \), that is,

\[
EH\varepsilon = EH\omega_j = E\omega_j\varepsilon = EX_n\varepsilon = EQ\varepsilon = EQ\omega_j = 0
\]

for all \( j \) and \( k \neq j \) and for all components \( X_n \) of \( X \). \hspace{1cm} (A.6)

### A.1. Preliminary Calculations

In this section we will gather a number of results that will be useful below.

1. Assume for the time being that \( H \) can be observed and consider the following ("forward" and "backward") regressions of \( H \) on \( X \),

\[
H = X\phi + u_2 \quad \text{(A.7a)}
\]

\[
X_n = \mu_n H + u_{n3} \quad \text{for } n = 1, \ldots, N, \quad \text{(A.7b)}
\]

where the disturbances \( u_2 \) and \( u_{n3} \) are assumed to satisfy the assumptions required for OLS to yield consistent estimates. It is easy to show that the probability limits of the OLS estimators of \( \phi \) and \( \mu = (\mu_1, \mu_2, \ldots) \) (which will be equal to the true parameter values) will be given by

\[
\phi = (EX'X)^{-1}EX'H \quad \text{and} \quad (A.8)
\]

\[
\mu = \frac{1}{EH^2}EX'H. \quad \text{(A.9)}
\]

The plim of the \( R^2 \) of equation (A.7a) will be given by

\[
ER^2_H = \text{plim} \ R^2(H|X) = \text{plim} \frac{\text{Explained SS}}{\text{Total SS}} = \frac{E(X\phi)'(X\phi)}{EH'H} \quad \text{(A.10)}
\]

Using (A.8) in the numerator of this expression, we have

\[
\phi'(EX'X)\phi = [(EX'X)^{-1}EX'H](EX'X)^{-1}EX'H = EH'X[(EX'X)^{-1}EX'H = EH'X(EX'X)^{-1}EX'H,
\]

(where we have made use of the fact that \( EX'X \) is a symmetric matrix) and therefore

\[
ER^2_H = \frac{EH'X(EX'X)^{-1}EX'H}{EH'H}. \quad \text{(A.10)}
\]
Using (A.8) and (A.9), this becomes
\[ ER^2_H = \frac{E H' X (E X' X)^{-1} E X' H}{E H' H} = \frac{E H' X \phi}{E H' H} = \mu' \phi. \quad \text{(A.10')} \]

ii. Assumptions (A.5) and (A.6) above imply
\[ E \varepsilon'_j \varepsilon_k = E (\rho_j \varepsilon + \omega_j)(\rho_k \varepsilon + \omega_k) = \rho_j \rho_k E \varepsilon^2, \quad \text{(A.11)} \]
\[ E \varepsilon'_j \varepsilon_j = E (\rho_j \varepsilon + \omega_j)(\rho_j \varepsilon + \omega_j) = \rho_j^2 E \varepsilon^2 + E \omega_j^2, \quad \text{(A.12)} \]
\[ E \varepsilon'_j H = E (\rho_j \varepsilon + \omega_j)H = 0, \quad \text{(A.13)} \]
\[ E \varepsilon'_j X = E (\rho_j \varepsilon + \omega_j)X = 0. \quad \text{(A.14)} \]

Using these results, we have
\[ EP'_j P_k = E (H' + \varepsilon'_j)(H + \varepsilon_k) = EH' H + EH' \varepsilon_k + E \varepsilon'_j H + E \varepsilon'_j \varepsilon_k \]
\[ = EH' H + \rho_j \rho_k E \varepsilon^2, \quad \text{(A.15)} \]
\[ EP'_j P_j = E (H' + \varepsilon'_j)(H + \varepsilon_j) = EH' H + EH' \varepsilon_j + E \varepsilon'_j H + E \varepsilon'_j \varepsilon_j \]
\[ = EH' H + \rho_j^2 E \varepsilon^2 + E \omega_j^2, \quad \text{(A.16)} \]
\[ EP'_j X_n = E (H' + \varepsilon'_j)X_n = EH'X_n + E \varepsilon'_j X_n = EH'X_n. \quad \text{(A.17)} \]
\[ EP'_j X = E (H' + \varepsilon'_j)X = EH' X + E \varepsilon'_j X = EH' X, \quad \text{(A.18)} \]
\[ EP'_j Q = E (H + \varepsilon'_j)Q = EH' Q + E \varepsilon'_j Q \]
\[ = EH' Q + E (\omega_j + \rho_j \varepsilon)Q = EH' Q. \quad \text{(A.19)} \]

To rewrite some of these expressions in a way that will be convenient, we define
\[ e_j \equiv \rho_j \sqrt{\frac{E \varepsilon^2}{E H^2}} \quad \text{and} \quad \text{(A.20)} \]
\[ U_j^2 \equiv \frac{E \omega_j^2}{E H^2}. \quad \text{(A.21)} \]

Factoring out $EH^2 (= EH' H)$ in (A.15) and (A.16), we have
\[ EP'_j P_k = EH' H + \rho_j \rho_k E \varepsilon^2 = EH^2 (1 + e_k e_k) \quad \text{(A.22)} \]
and
\[ EP'_j P_j = EH' H + \rho_j^2 E \varepsilon^2 + E \omega_j^2 = EH^2 (1 + e_j^2 + U_j^2) \quad \text{(A.23)} \]
Let us define $ER^2_j$ as the probability limit of the coefficient of determination of a regression of $P_j$ on the vector $X$, $R^2(P_j|X)$. By analogy with (A.10), we have

$$ER^2_j = \text{plim} \ R^2(P_j|X) = \frac{EP_j'X(XX')^{-1}EX'P_j}{EP_j'P_j}. \quad (A.24)$$

Notice that, using (A.18), the numerator of this expression can be written

$$EP_j'X(XX')^{-1}EX'P_j = EH'X(XX')^{-1}EX'H.$$

Substituting this expression into (A.24), and using (A.4) and (A.10), we have

$$ER^2_j = \frac{EH'X(XX')^{-1}EX'H \cdot EH'H}{EP_j'P_j} = ER^2_{H^r}. \quad (A.25)$$

### A.2. Measurement Error Bias

Consider now what happens when we estimate equation (A.1) using an imperfect proxy $P_j$ for the stock of human capital, $H$. By analogy with (A.2), the probability limit of the resulting OLS estimator, $\hat{\beta}_j$, is given by

$$\text{plim} \ \hat{\beta}_j = \frac{EP_j'Q - EP_j'X(XX')^{-1}EX'Q}{EP_j'P_j - EP_j'X(XX')^{-1}EX'P_j} \times \frac{EH'H - EH'X(XX')^{-1}EX'H}{EP_j'P_j} = A^*B. \quad (A.26)$$

We will now consider in turn each of the two factors in the last expression. Using (A.18), (A.19) and (A.2) in the first term, we have

$$A = \frac{EP_j'Q - EP_j'X(XX')^{-1}EX'Q}{EH'H - EH'X(XX')^{-1}EX'H} = \frac{EH'Q - EH'X(XX')^{-1}EX'Q}{EH'H - EH'X(XX')^{-1}EX'H} = \text{plim} \ \hat{\beta}_H = \beta.$$
for $E R_j^2$ given in (A.25), and (A.26) to obtain

$$B = \frac{E H' H - E H' X (E X' X)^{-1} E X' H}{E P_j' P_j - E P_j' X (E X' X)^{-1} E X' P_j}$$

$$= \frac{E H' H - E H' X (E X' X)^{-1} E X' H}{1 - \frac{E P_j' X (E X' X)^{-1} E X' P_j}{E P_j' P_j}}$$

$$= \frac{r_j - \frac{E H' X (E X' X)^{-1} E X' H}{E H' P_j' P_j} \frac{E H' P_j' P_j}{1 - E R_j^2}}{1 - E R_j^2}$$

$$= \frac{r_j(1 - E R_j^2)}{1 - E R_j^2}.$$

Collecting results, we arrive at the following formula, which shows the attenuation effect as a function of $P_j$’s reliability ratio, $r_j$, and $E R_j^2$:

$$\text{plim} \hat{p_j} = \frac{r_j(1 - E R_j^2)}{1 - E R_j^2} \beta \equiv a_j \beta,$$  \hspace{1cm} (A.27)

where $a_j$ is the attenuation coefficient for series $P_j$.\(^{18}\)

### A.3. Estimating $r_j$ and $E R_j^2$

Equation (A.27) can be used to obtain a meta-estimate of $\beta$ that will be clean of measurement error bias. For this, we need consistent estimates of $r_j$ and $E R_j^2$ or, equivalently (by equation (A.10')) of $r_j$ and of the vectors $\mu$ and $\phi$.

The two-stage procedure we use to estimate $r_j$ and $\mu$ is described in Section 5.1 of the text. Here we calculate the probability limits of the preliminary (“pairwise”) estimates of these parameters, $\hat{r}_{jk}$ and $\hat{\mu}_{jn}$, that are obtained by estimating the first-stage regressions

$$P_k = r_{jk} P_j + u_{jk} \quad \text{for } k = 1, \ldots, J \text{ with } k \neq j$$  \hspace{1cm} (A.28)

and

$$X_n = \mu_{jn} P_j + u_{jn} \quad \text{for } n = 1, \ldots, N$$  \hspace{1cm} (A.29)

and serve as data in the second-stage regressions.

\(^{18}\) Notice that if $\beta_H$ is biased, say because of a non-zero correlation between $H$ and $u_1$ in (1), then we have $\text{plim} \hat{p_j} = a_j (\beta + \text{bias})$. Hence, testing for a non-zero constant in equation (24) in the text will not help us rule out other sources of bias.
Using (A.22) and (A.4) in the expression for the OLS estimator of $r_{jk}$ in (A.28), we have

$$\text{plim} \  \hat{r}_{jk} = \frac{EP_j' P_k}{EP_j^2} = \frac{EH^2(1 + e_j e_k)}{EP_j^2} = r_j (1 + e_j e_k).$$
(A.30)

Similarly, for equation (A.29), equation (A.9) with $P_j$ replacing $H$ implies, using (A.17), (A.4), and (A.9), that

$$\text{plim} \  \hat{\mu}_{jn} = \frac{EP_j' X_n}{EP_j^2} = \frac{EH^2}{EP_j^2} EP_j^2 = r_j \mu_n.$$
(A.31)

Finally, consider any reverse regression of the form

$$P_j = X\phi + u_{xj}.$$  \hspace{1cm}  (A.32)

Then, equation (A.8) with $P_j$ replacing $H$ yields the following plim for the corresponding OLS estimate, $\hat{\phi}_{j}$,

$$\text{plim} \  \hat{\phi}_{j} = (EX'X)^{-1}EX'P_j = (EX'X)^{-1}EX'H = \phi,$$
(A.33)

where we have made use of equations (A.17) and (A.8) again. Hence, any regression of the form (A.32) will yield a consistent estimate of $\phi$, and so will the restricted system of all such regressions we use to estimate this parameter vector (equation (26) in the text), because the resulting estimator will be a weighted average of asymptotically unbiased ones.

References


This article has been cited by: