

Measurement error in growth regressions: re-evaluating the role of education*

Miguel Portela[†] Rob Alessie[‡] Coen Teulings[§]

19th November 2003

Abstract

This paper analyses the measurement error in education data that is created by the use of the perpetual inventory method. It is shown that there is a systematic difference between census or survey data and constructed data for education. Moreover, this systematic difference interferes with the estimated impact of education on economic growth. Procedures that reduce the impact bias in growth regressions are proposed. Once we control for measurement error, the results indicate that both the change in education and its level are relevant for economic growth.

1 Introduction

Several studies have analyzed the impact of education on economic growth. However, the validity of the results depend dramatically on the reliability of education data. Authors like Benhabib and Spiegel (1994), using Kyriacou (1991) data, and Barro and Sala-i-Martin (1999), using Barro and Lee (1996) data, conclude that it is the level of education, not its change, that has an impact on economic growth. As such, they find empirical evidence in favor of Nelson and Phelps' (1966) argument that growth is driven by the stock of human capital through its effect on countries' capacity to innovate or catch up, and refute Lucas' (1988) conclusion that it is the accumulation of human capital that drives growth. Theoretically, their empirical evidence lends support to the human capital type of endogenous growth models. Krueger and Lindahl (2001), using both data sets, argue that these results are highly influenced by the measurement error in the average education of countries. They conclude that "both the change and initial level of education are positively correlated with economic growth", finding empirical evidence in favor of

*The first author gratefully acknowledges financial support by the Portuguese Foundation of Science and Technology (FCT; ref. SFRH/BD/5114/2001). We would like to thank Thijs van Rens, and the participants of the Tinbergen Institute Seminar, for helpful comments.

[†]Minho University and Tinbergen Institute

[‡]Utrecht School of Economics

[§]Erasmus University Rotterdam and Tinbergen Institute

Lucas' argument. These contradictory results highlight the relevance of the quality of education data in the growth literature.

The aim of this paper is, first of all, to analyze the systematic difference between census and constructed data points in Barro and Lee (2001) data set, and, secondly, to propose a simple method to improve the estimation of the impact of education on growth using knowledge on the systematic difference between census and non-census data. We apply our procedure to these data, given that it provides information on a broader number of countries than most other data sets, being moreover the most commonly used source of information on education in empirical studies of growth. It should be kept in mind that this is just one further tentative to overcome our inability to observe the true education in each country.

This work diverges from the literature dealing with the difference between education and human capital, and their role in economic growth.¹ The goal here is not to argue that education is not the right measure of human capital, and to propose an alternative measure of human capital, or to address the issue of quality differences in education systems across countries. Instead, we are mainly concerned with the impact of the measurement error in education in the conclusions on the relation between education and growth. The natural extension of our work would be to include quality measures, and other measures of human capital, like the ones referred in Hanushek and Kimko (2000).

2 Sources of data on education

In recent years the most commonly used data set on international education attainment is the one released by Barro and Lee. Alternative sources of information are Kyriacou (1991), de la Fuente and Doménech (2002), and Cohen and Soto (2001). These data sources have in common the fact that they rely on census data when available. This is one important advantage over the data released by Nehru *et al.* (1995), which ignores census data on attainment levels. De la Fuente and Doménech (2002) criticise this choice, and argue that it is difficult to justify “discarding the only direct information available on the variables of interest.”

Kyriacou's data applies to the labour force, and was estimated for the period 1965 to 1985 at five year intervals. In order to estimate the average school years for each country, Kyriacou assumes that the relationship between average school years in the labor force and the enrollment ratios in primary, secondary and higher education is relatively constant over time and across countries.

Barro and Lee (1993) build their data on educational attainment from census or survey data. When this information is not available, the authors use a perpetual inventory method based on enrollment data in order to generate either a forward-flow, or a backward-flow. The flows are constructed from the benchmark stocks defined in the census or survey data. When feasible, the estimation of the missing values on attainment levels is a weighted average of the forward-flow and the interpolation between two benchmarks. If the interpolation is not available, the fill-in procedure implies that the missing values are determined either by the forward-flow, or by the backward-flow.

¹See, for example, Hanushek and Kimko (2000), and Woessmann (2000).

The data available in Barro and Lee (2001) improves the previous two versions of the data released in Barro and Lee (1993) and Barro and Lee (1996). In the new data set, the fill-in procedure for missing census/survey observations uses gross enrolment rates adjusted for repeaters. In the construction of average years of schooling, it also adjusts for changes in the duration of years of schooling. This data set gives information on estimated educational attainment for the population over age 15 and over age 25 at five-year intervals between 1960 and 2000. “In principle, Barro and Lee’s procedure should be superior to Kyriacou’s because it makes use of more information and does not rely on such strong implicit assumptions” (de la Fuente and Doménech, 2002).

Although Barro and Lee’s data improve on the quality of the alternative sources, they still received some criticism. De la Fuente and Doménech (2002) construct a “revised version of the Barro and Lee (1996) data set for a sample of OECD countries using previously unexplored sources and following a heuristic approach to obtain plausible time profiles for attainment levels by removing sharp breaks in the data that seem to reflect changes in classification criteria” (de la Fuente and Doménech, 2002).²

De la Fuente and Doménech’s data applies to the fraction of the population 25 and older, and it is only available for a set of 21 OECD countries. The authors’ goal is to reconstruct a plausible schooling profile for each country, and to “avoid unreasonable jumps in the series by choosing the most plausible figure when several are available for the same year, and by reinterpreting some of the data” (de la Fuente and Doménech, 2002). In order to fill the missing observations on schooling attainments they interpolate when possible, and apply backward or forward projections in the other situations. The authors “avoided the use of flow estimates based on enrollment data because they seem to produce implausible time profiles” (de la Fuente and Doménech, 2002). The expression “reasonable guesses” used in the paper seems to translate one important idea behind the construction of the data set. The authors state that “the construction of our series involves a fair amount of guesswork,” and that their data “look more plausible than most existing series, at least in terms of their time profile.”

Cohen and Soto (2001) extend the work of de la Fuente and Doménech (2002) to several other countries. An important difference to de la Fuente and Doménech is that Cohen and Soto allow for the use of enrolment data when needed. The authors have constructed a data set for 95 countries with information on education achievement from 1960 to 2000, for ten year interval, plus a projection for 2010. Their methodology is to “minimize the extrapolations and keep the data as close as possible to those directly available from national censuses” (Cohen and Soto, 2001). They argue that some of the differences between their data and the data provided by Barro and Lee (2001) can be explained by: (i) divergences in classification; (ii) the use of more census information than Barro and Lee; (iii) the use of a different methodology for extrapolating the missing data; (iv) errors in Barro and Lee data.

At this stage, the conclusion would be that, in spite of the improvements in data, so far measurement error in average education across countries remains a problem. Barro and Lee data is highly volatile, and presents frequent decreases over time within countries.

²These two data sets are not directly comparable since Barro and Lee’s data is based on people having completed some educational level, while de la Fuente and Doménech’s data applies to people who have attended some educational level.

De la Fuente and Doménech argue in favor of their data, but one of its problems is that it is only available for a sample of 21 countries. Also, they do not solve entirely the problem of measurement error, although they present the best reliability results³ when compared to the other sources of information on countries education. Cohen and Soto's data is only available on 10-year intervals, and for a sample of 95 countries, and they are also subject to the criticism that measurement error problems were not solved. Kyriacou data is very problematic given the estimation procedure used, and it is only available for the period 1965–1985 in a five year interval.

3 How systematic is the difference between census and non-census data?

We will now analyse how systematic is the difference between census/survey data and non-census data in Barro and Lee's data set. Contrary to de la Fuente and Doménech's (2002, p. 25) argument that there is no reason to suspect that the available education data contains systematic biases, we believe that there is a systematic error in the education data released by Barro and Lee. If true, this systematicity could lead to the overestimation of the coefficient of education in a growth regression.⁴

Barro and Lee's fill-in procedure for missing values on education is based on interpolation, when feasible, and on forward flow if available. Backward flow is used when none of the previous values is available. This procedure implies that missing values are estimated differently according to the type of observation: (i) observations before the first census/survey within a country (type A); (ii) observations between census data within countries (type B); (iii) observations after the last census within each country (type C). Our empirical strategy to test for systematic difference between census and non-census data will take into account for these differences.

The hypothesis to test is that the perpetual inventory method used to impute values for education in Barro and Lee data underestimates the values of education. If this is true, we should observe in the data that: (i) the variation of education between two consecutive census observations should be higher than the same variation between non-census observations, and (ii) the variation between a non-census to a census point should be the highest among the three variations.

The underestimation can result from the assumption that the survival rate used in the original estimations for the missing values is independent of the educational level. In their own words, Barro and Lee (1993, p. 374) state that "some error is introduced (...) if educational attainment is growing rapidly, because the older people then have less human capital and a greater probability of dying." If average education within a country is rising,

³See de la Fuente and Doménech (2002).

⁴Suppose our model is defined as $y_t = \beta_1 + \beta_2 x_t^* + \varepsilon_t$, where the observed value of x is $x_t = x_t^* + \gamma t + u_t$, and x is underestimated at a constant rate γ over time. The model that will be estimated based on the observed variables is $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t - \beta_2 \gamma t - \beta_2 u_t$. A general result from error in variables models is that the inconsistency in the estimation of β_2 is given by $\frac{\text{cov}(x_t, \varepsilon_t - \beta_2 \gamma t - \beta_2 u_t)}{\text{var}(x_t)} = \frac{-\beta_2 \gamma \sigma_t - \beta_2 \sigma_u}{\sigma_x}$. Given that γ is negative by definition, the traditional downward bias will be smaller, and it may even be an upward bias.

as it seems to be the case for an important portion of the countries, the implication would be an underestimation of the educational attainment. The increase in the schooling level of a population occurs mainly because the younger generations are more educated. In this case, the estimation procedure underestimates the survival of more educated individuals, resulting in a lower attainment for the country as a whole. The same idea is identified in Barro and Lee (2001, p. 545), when they say that “in a typical country in which educational attainment is growing, mortality would be higher for the older people who are less educated. Then the assumption of uniform mortality can cause a downward bias in the estimation of the total educational stock.”

Four variables will be built in order to identify the systematic error in the education variable and to propose a procedure that handles the problem. The variable *Before* measures the lag till the first census and it applies to observations of type A. The variable *Last* records the number of periods that have elapsed since the previous census for non-census observations for type B observations. If the observation was itself collected in a census, then a third variable, *LastC*, equals *Last*, being zero otherwise. The variable *After* measures the lag till the last census, and is valid for observations of type C. Each one of these variables assumes a non-zero value for its own type, and zero for the other types.

These variables focus on the timing of each observation on education, with respect to the timing of the census. Our argument is that we always have to consider the period that has elapsed to the census observation, since errors accumulate over time. If, alternatively to our four variables, we use only a dummy for census observation, we will be capturing the average change between a non-census and a census observation. This is the average correction for the accumulated errors over time, not the underestimation of education each period.

Figure 1 shows the argument. Imagine an hypothetical country with 9 observations. In the horizontal axis we have the time dimension, while the vertical axis plots the average education level in each period. The steeper and darker line represents the evolution of the true education. To render the explanation simpler, assume that the true education follows a constant trend. The observations represented by an empty square, located in this line, represent the census information available. The circle dots represent the estimated points using the enrollment data and the benchmark census information. We also assume that the estimation process leads to a constant trend, that underestimates the true value. This is represented by the lighter line. The filled square dots represent the values of education that would be estimated for periods in which we have census data.

The change in education from period 4 (the empty circle in period 4) to period 5 (the empty square in period 5) can be decomposed as the variation predicted by the perpetual inventory method (the filled square dot over the lighter line in period 5), plus the accumulated errors since period 2, originated by the underestimation. The jump between the non-census (hypothetical) and census data points in period 5 (the difference between the filled square dot and the empty square dot) is proportional to the time elapsed since the previous census. In period 3 the error is given by the distance between the empty circle and the steeper line. In period 4 the difference between the empty circle and the steeper line gives the accumulated error in period 3 and 4. The error specific to period 4 can be retrieved if we imagine a non-census trend line departing from the true education

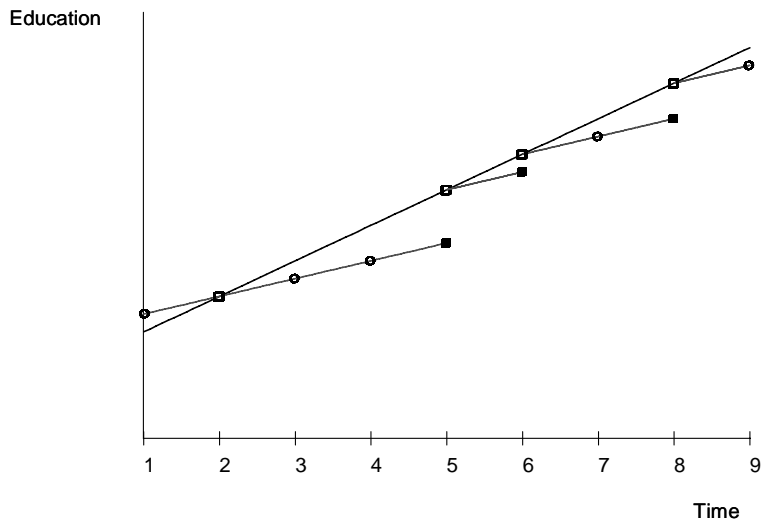


Figure 1: Plot of education with census and non-census data

value in period 3.

If we use a dummy for census observation, the variable assumes the value zero in period 4 and one in period 5. Its coefficient would be a weighted average of the changes associated with different lags till the previous census. Saying it in another way, it captures the average change between a non-census and a census observation, which implies that it overestimates the difference between census and non-census observations in each period.

In a regression we can imagine a general trend originated by the combination of census and non-census data. Both trends depicted in the graph diverge from this one. If the above hypothesis is true, the trend associated with the census observations should be steeper than the one underlying the non-census data. The variable *LastC* represents the difference between the trend within census data and the general trend, while the variable *Last* represents the divergence between the trend estimated with non-census information and the same general trend.

In reality, the estimation procedure implemented by Barro and Lee is more complex, which implies that we need additional variables in our model, *Before* and *After*. As explained before, the estimation is different according to the relative position of each observation in the distribution of the census data points within countries. This is the reason why we have to consider that a backward prediction, period 1 in our graph, may differ from a forward prediction. We also have to take into account possible differences between forward prediction after the most recent census available, period 9, and forward prediction between census data points. Looking at period 1 in the graph one would expect that the coefficient on the backward distance till the first census should be positive. The reason is that the forward underestimation translates into an overestimation in a backward prediction.

3.1 Data description

Data on education (*Edu*), and its source, is taken from Barro and Lee (2001). We will focus our attention on population aged 15 and over. The data source is described by the variable *Census*, which is a dummy variable that assumes the value 1 if the information on the educational attainment was based on a census or survey, and 0 if the education level was estimated. The variables *Before*, *Last*, *LastC*, and *After* are constructed from *Census* variable as described above. All the variables are available on five year intervals, between 1960 and 2000.

Tables 6, 7, 8, and 9 in the appendix provide a description of the data used in the analysis of education. Our sample is composed by 985 observations corresponding to 116 countries. Only 32% of the information on education is based on census/survey data (Table 6). Among the constructed census variables, *Before* is the one with higher incidence of zeros, and its values range between 0 and 5. Out of 9 possible periods, 46% of the countries in the sample have 2 or less education census points, 28% have 3, and only 26% of the countries have 4 or more census observations on education (Table 7). The distribution of countries per period is relatively balanced, with the minimum number of observations in a given period of 104. From Table 8 we can also observe an increase of countries' average education over time. When we take the sub-sample of census observations, Table 9, there is a clear unbalanced distribution of observations per period, with 1980 being the year with more census observations, 67 out of 111. In this case, both, the average education, and its increase, are higher, when compared with the full sample.

3.2 Empirical evidence

We will now proceed with the empirical test of our hypothesis. First, we define a model in levels, taking into account for time and country specificities,

$$Edu_{it} = \gamma_t + \beta_1 Before_{it} + \beta_2 Last_{it} + \beta_3 LastC_{it} + \beta_4 After_{it} + \eta_i + \varepsilon_{it} \quad (1)$$

where Edu_{it} is the education level of country i , in period t , γ_t is the period t specific effect, η_i is country i 's specific effect, and ε_{it} is a white noise error term. Second, we take first-differences of equation (1),

$$\Delta Edu_{it} = \gamma_t + \beta_1 \Delta Before_{it} + \beta_2 \Delta Last_{it} + \beta_3 \Delta LastC_{it} + \beta_4 \Delta After_{it} + \Delta \varepsilon_{it} \quad (2)$$

where $\Delta Edu_{it} = Edu_{it} - Edu_{i,t-1}$, and the same transformation applies to the other variables. Finally, we depart from equation (1), and define a dynamic model for education and its relationship with the census variables,

$$Edu_{it} = \gamma_t + \alpha Edu_{i,t-1} + \beta_1 Before_{it} + \beta_2 Last_{it} + \beta_3 LastC_{it} + \beta_4 After_{it} + \eta_i + \varepsilon_{it} \quad (3)$$

The results for the estimation of equations (1), (2) and (3), are presented in Table 1. As a general remark, all the regressions are globally significant, and the coefficients on the time dummies are jointly statistically different from zero. We used 116 countries in order to implement our estimations, and the number of observations varied between 753 for the dynamic model estimated by the GMM procedure suggested by Arellano and Bond (1991), *AB*(1991), and 985 for the model in levels, *FixedEffects* column.

Table 1: Education Regressions

Variable	Levels		First-differences		Dynamic model	
	Fixed Effects	OLS	Fixed Effects	AB(1991)	BB(1998)	
LagEdu				0.369** (0.099)	0.923** (0.025)	
Before	0.391** (0.072)	0.250** (0.055)	0.140* (0.065)	0.167* (0.080)	-0.039 (0.027)	
Last	-0.200** (0.032)	-0.198** (0.027)	-0.186** (0.031)	-0.108** (0.028)	0.031 (0.025)	
LastC	0.199** (0.032)	0.202** (0.029)	0.193** (0.029)	0.210** (0.031)	0.191** (0.031)	
After	-0.214** (0.057)	-0.272** (0.056)	-0.316** (0.085)	-0.153* (0.069)	-0.054* (0.026)	
Wald joint	83.008**	70.190**	57.522**	140.845**	2374.090**	
Wald time	610.577**	566.373**	39.747**	50.214**	22.791**	
F-Test: all ui=0	251.61**		1.09			
Sargan				75.844	98.022	
Sargan-df				73	112	
SarganDiff					22.179	
AR(1)	5.849**	-1.070	-4.611**	-3.021**	-5.202**	
AR(2)	2.594**	-0.366	-1.579	-0.548	-0.474	
Nobs	985	869	869	753	869	
Ncountries	116	116	116	116	116	

Significance levels : † : 10% * : 5% ** : 1%. (Standard errors in parentheses)

The dependent variable is Education. All regressions include time dummies.

The instruments are

- (i) first-difference equations: level of education lagged two periods and earlier;
- (ii) level equations: first-difference of education lagged one period.

The variables Before, Last, LastC and After are taken as exogenous.

The coefficient on the variable *LastC* is our focus of interest. Over the five regressions it remains relatively stable, varying between 0.193 and 0.210, economically significant and always statistically significant. We observe that, on average, the underestimation induced by the perpetual inventory method is around a fifth of a school year per period of five years. This leads to a significant spurious variation in education both, over time, and within countries. This feature is prone to interfere with the results reached in the growth regressions, as will be seen in the next section.

In column 1 of Table 1 we report the results of the estimation of equation (1) using the fixed-effects estimator. The results in columns 2 and 3 refer to the model in first-differences defined in equation (2). In column 2 the estimation procedure is an OLS, while in column 3 we implement the Within estimator. In this second alternative we test for the possibility of country specificities in the changes in education. In the estimation of

equations (1) and (2) the coefficients on *Last* and *LastC* are identical, but with opposite signs. Remember that *LastC* is only nonzero for census observations. So, for census observations these coefficients cancel out. The interpretation is that, in the regression, the general trend is the one associated with the census observations, and what happens is that non-census observations fall short of their average true value in approximately 0.2 of a year every period. Once we achieve a census data point this measurement error is corrected. This correction is proportional to the number of periods since the previous, or the nearest, census observation.

In columns 1, 2, and 3, the coefficient of the variable *Before* is always positive and significant, although it is not stable across the different estimations. Its sign is justified by the fact that these observations were estimated by the backward flow. This supports our hypothesis. If the estimation of missing values underestimates the true value of education, it means that the general trend implicit in the variation between census observations is steeper than the one associated with non-census observations. In a backward prediction, the estimated value will be above the true value, since we are underestimating a decrease from the next census observation. The coefficient on the variable *After* is negative in this same regressions, which is also expected. In this case we conclude that the education level after the last census is also underestimated. As expected, we reject the hypothesis that all the specific effects are equal to zero. When we apply the fixed-effects estimator to the first-differences estimator we do not reject the hypothesis that all the specific effects are equal to zero. This way, the results in column 2 are the valid ones, when compared to column 3.

Finally, equations 4 and 5 report the results of the estimation of the dynamic model in first-differences as suggested by Arellano and Bond (1991), and a system estimation following Blundell and Bond (1998), respectively. In both regressions we take the census variables as exogenous. Although we do not reject the validity of the instruments used in column 4, the results found by Blundell and Bond suggest that the preferred estimation is the one presented in column 5, the system estimation. The Sargan difference test does not reject the validity of the extra instruments used in the system estimation, when compared with the first-differences estimation of the dynamic model. Also the Sargan test does not reject the instruments used in the system estimation. The tests for first-order and second order serial correlation in the error term perform accordingly to the assumptions underlying the dynamic estimation procedures (no serial correlation in the disturbances).

Using a dynamic formulation of the model and implementing the system estimation leads to a slightly smaller coefficient than the one obtained in the model in first-differences for the variable *LastC*, which achieves in the dynamic formulation of the model the value 0.191. In the dynamic model, for the system estimation, the coefficients of the variables *Before* and *Last* become statistically insignificant, and its absolute value decreases significantly. The variable *After* is still statistically significant, but the size of the coefficient is significantly smaller than the one obtained in the first-differences model. Our interpretation is that the coefficient on lag education captures the evolution over time of the non-census observations, implying that the effect of the census observations is identified by the coefficient of *LastC*.

3.3 How to correct for the systematic difference?

In the previous section we show how different census and non-census observations are. Now, the question is how to use this information to improve the quality of the data. First, we will use the results from the static model. One of the main criticism of de la Fuente and Doménech (2002) is that Barro and Lee data has sharp breaks in its series. With the following procedure we attenuate them, and reduce the serial correlation in the measurement error in education. The idea is to transform the original data constructed by Barro and Lee according to the following expression

$$PEdu_{it} = Edu_{it} - \beta_1 Before_{it} - \beta_2 Last_{it} - \beta_3 LastC_{it} - \beta_4 After_{it} \quad (4)$$

where $PEdu_{it}$ is the predicted education after compensating for the systematic error. Using the results in column 2 of Table 1, the predicted value of education would be given by

$$PEduFD_{it} = Edu_{it} - 0.250 * Before_{it} + 0.198 * Last_{it} - 0.202 * LastC_{it} + 0.272 * After_{it} \quad (5)$$

Second, we can use the results for the dynamic model to correct for the systematic measurement error. Using the results from columns 4 and 5 of Table 1, the predicted value of education would be given by

$$PEduAB_{it} = Edu_{it} - 0.167 * Before_{it} + 0.108 * Last_{it} - 0.210 * LastC_{it} + 0.153 * After_{it} \quad (6)$$

and

$$PEduBB_{it} = Edu_{it} + 0.039 * Before_{it} - 0.031 * Last_{it} - 0.191 * LastC_{it} + 0.054 * After_{it} \quad (7)$$

We also define two alternative predictions based on equations (6) and (7), but in which the transformation is not applied to census observations. These extra transformations are identified by $PEduAB2$ and $PEduBB2$, respectively.

In order to evaluate these new measures we calculate the correlations among the different education variables, including the alternative sources of information, $EduCS$ the data released by Cohen and Soto (2001), $EduDD$ the data constructed by de la Fuente and Doménech (2002), and also Kyriacou (1991)'s data, $EduKY$. The correlations of the average years of schooling are given in Tables 2 and 3. In the first case the correlations are evaluated for the variables in levels, while in the second the correlations are computed for the variables in first-differences.

The correlation between Barro and Lee education level and the predicted education variables is very high. The correlation of these variables and the alternatives is lower, but it remains high, in particular with $EduCS$. This correlation is always higher than 0.88, the correlation between $PEduBB$ and $EduKY$. The lowest correlation occurs between Kyriacou's data and de la Fuente and Doménech, 0.533. When we calculate the correlations using the variables in first-differences, Table 3, the correlations drop. Among Barro and Lee education and the predicted variables the correlation is still very high. However, when we compare with the alternative sources of information, the correlations drops significantly. Again the correlations are higher with Coen and Soto data. The correlation

between differenced Barro and Lee data and differenced Coen and Soto data is approximately 0.2. The correlations for the predicted variables are similar, with a smaller value for the prediction using the first-differences estimation, $PEduFD$.

The reliability ratios for the variables of interest, Edu , $PEduFD$, $PEduAB$, $PEduAB2$, $PEduBB$ and $PEduBB2$, are calculated using the alternative measures of education, $EduCS$, $EDuDD$, and $EduKY$. According to Krueger and Lindahl (2001) the reliability ratio can be “derived by regressing one measure of years of schooling on the other.” The corresponding reliability ratios are reported in Table 4. This Table also reports the mean and the variance of the variables of interest for each one of the samples and transformations used in the estimation of the reliability ratios. We implemented the regressions in levels and in first-differences. In the second case we distinguished between the 5, the 10, and the 20 year interval data. In all regressions we included time dummies, and for the level regressions we controlled for country specific effects. The level regressions are estimated by the fixed effects estimator, while the first-differences estimations are implemented by OLS.

Table 2: Correlations among Education Measures in Levels

	Edu	PEduFD	PEduAB	PEduAB2	PEduBB	PEduBB2	EduCS	EduDD	EduKY
Edu	1 (985)								
PEduFD	0.987 (985)	1 (985)							
PEduAB	0.994 (985)	0.998 (985)	1 (985)						
PEduAB2	0.995 (985)	0.998 (985)	1.000 (985)	1 (985)					
PEduBB	0.997 (985)	0.990 (985)	0.997 (985)	0.995 (985)	1 (985)				
PEduBB2	1.000 (985)	0.990 (985)	0.996 (985)	0.997 (985)	0.998 (985)	1 (985)			
EduCS	0.956 (420)	0.956 (420)	0.960 (420)	0.960 (420)	0.953 (420)	0.956 (420)	1 (420)		
EduDD	0.892 (155)	0.888 (155)	0.893 (155)	0.893 (155)	0.892 (155)	0.893 (155)	0.933 (80)	1 (155)	
EduKY	0.886 (420)	0.896 (420)	0.892 (420)	0.893 (420)	0.880 (420)	0.885 (420)	0.910 (143)	0.533 (90)	1 (420)

Number of observations in parentheses.

Table 3: Correlations among Education Measures in First-Differences

	DEdu	DPEduFD	DPEduAB	DPEduAB2	DPEduBB	DPEduBB2	DEduCS	DEduDD	DEduKY
Dedu	1 (869)								
DPEduFD	0.883 (869)	1 (869)							
DPEduAB	0.845 (869)	0.974 (869)	1 (869)						
DPEduAB2	0.968 (869)	0.972 (869)	0.937 (869)	1 (869)					
DPEduBB	0.766 (869)	0.886 (869)	0.965 (869)	0.847 (869)	1 (869)				
DPEduBB2	0.996 (869)	0.875 (869)	0.834 (869)	0.961 (869)	0.763 (869)	1 (869)			
DEduCS	0.369 (335)	0.348 (335)	0.355 (335)	0.364 (335)	0.362 (335)	0.381 (335)	1 (335)		
DEduDD	0.068 (135)	0.020 (135)	0.048 (135)	0.036 (135)	0.110 (135)	0.084 (135)	0.391 (60)	1 (135)	
DEduKY	0.052 (324)	0.052 (324)	0.047 (324)	0.052 (324)	0.047 (324)	0.056 (324)	0.201 (66)	0.190 (70)	1 (324)

Number of observations in parentheses.

With the exception of *PEduBB*, the predicted variables have higher means and variances. The higher dispersion is determined by the *After* and *Before* type of predictions. The reliability ratios indicate that the predicted measures of education are more reliable than *Edu*. When we use *EDuCS* as the comparison variable, the reliability ratio for *Edu* is 0.39, while for *PEduFD* is 0.43. The fact that the predicted variable has an higher variance reinforces this conclusion. Using *EduKY* as the comparison variable yields similar results. The regressions using *EduDD* as the dependent variable have very small and insignificant reliability ratios. One possible explanation is the fact that in this case we are using a particular sample of 20 OECD countries. Krueger and Lindahl (2001) note that if the errors in the variables used in each regression “are positively correlated, the estimated reliability ratios will be biased upward.”

Table 4: Reliability Ratios and Descriptive Statistics

Variables	Edu	PEduFD	PEduAB	PEduAB2	PEduBB	PEduBB2
	Levels					
Mean	5.028	5.280	5.124	5.163	4.995	5.080
Variance	8.299	8.881	8.551	8.592	8.255	8.337
EduCS	0.388**	0.427**	0.428**	0.432**	0.393**	0.397**
EduDD	0.026	-0.004	0.005	0.007	0.036	0.039
EduKY	0.336**	0.382**	0.370**	0.381**	0.332**	0.339**
	First-Differences, 5 Year Data					
Mean	0.350	0.501	0.438	0.439	0.367	0.368
Variance	0.192	0.147	0.156	0.157	0.193	0.200
EduCS	0.141**	0.142**	0.139**	0.148**	0.133**	0.145**
EduKY	0.063	0.068	0.057	0.068	0.054	0.069
	First-Differences, 10 Year Data					
Mean	0.697	0.989	0.869	0.869	0.731	0.731
Variance	0.370	0.296	0.312	0.304	0.403	0.384
EduCS	0.257**	0.279**	0.279**	0.281**	0.264**	0.262**
EduKY	0.309*	0.341*	0.327*	0.343*	0.289*	0.312*
	First-Differences, 20 Year Data					
Mean	1.411	2.003	1.756	1.758	1.474	1.480
Variance	0.723	0.622	0.642	0.616	0.826	0.744
EduCS	0.401**	0.395**	0.392**	0.420**	0.368**	0.407**
EduCS+	0.414**	0.478**	0.452**	0.481**	0.375**	0.434**
EduKY	0.761**	0.946**	0.931**	0.933**	0.737**	0.734**

Significance levels: † : 10% * : 5% ** : 1%.

The reliability ratio is the coefficient in the regression of the alternative measure of education using the Variables in columns as regressors. All regressions include time dummies. The model in levels include country specific effects. EduCS+ estimations include only the two most recent periods.

If we use 5 year differenced data the signal to noise ratio is reduced, and the reliability ratios drop significantly. The reliability of the variables of interest is similar. For the variables *PEduFD*, *PEduAB*, and *PEduAB2*, the variance of their changes is smaller than the variance of the changes in *Edu*. Using this transformation we are not able to assert if our transformations improves the quality of the data. When we use changes over longer periods the reliability of the data increases, and the differences of the reliability of the predicted measures of education when compared with the original data are more clear. Using differences between observations 20 years apart shows clearly the improvement in the reliability of the data when we correct for its source. A general conclusion would be that *PEduFD* is, on average, the most reliable correction of the data.

4 Growth regressions: what changes?

Having analyzed the difference in education data according to its source, we will now re-evaluate the role of education in economic growth. We will estimate the macro-Mincerian growth equation using the different measures of education. Our empirical growth model is defined as

$$\text{LogGDP}_{it} = \gamma_t + \alpha \text{LogGDP}_{i,t-1} + \beta_1 S_{it} + \beta_2 S_{i,t-1} + \tau X_{it} + \eta_i + v_{it} \quad (8)$$

where LogGDP_{it} is the logarithm of income per worker, S_{it} stands for the average education of country i in period t , η_i is country i 's specific effect, and v_{it} is the white noise error term. The vector X_{it} stands for the extra set of regressors K_i , investment share of GDP, and SF , State failure, which characterize a complete collapse of the central authority. Income and investment share are taken from the Penn World Tables, version 6.1, while the variable State failure is obtained from the Polity IV Project. In one of the specifications we will include the census variables defined above.

We can transform equation (8) as

$$\text{LogGDP}_{it} = \gamma_t + \alpha \text{LogGDP}_{i,t-1} + \beta_1 \Delta S_{it} + (\beta_2 + \beta_1) S_{i,t-1} + \tau X_{it} + \eta_i + v_{it} \quad (9)$$

in order to identify the cumulative effect of changes in education.

Given the existence of a country specific effect in equation (9), its estimation by OLS and by the usual panel models, fixed or random effects, is inconsistent. The reason is that, by definition, $\text{LogGDP}_{i,t-1}$ in equation (9) is always correlated with η_i . One possible solution to overcome this problem is to take first differences in equation (9) to eliminate the fixed effect. Arellano and Bond (1991) first-differenced generalized method of moments, and Blundell and Bond (1998) system estimation are two of the most applied solutions.

Using Arellano and Bond (1991) procedure avoids the bias introduced by omitted time-invariant variables, and allows, under certain circumstances, to obtain consistent estimates in the presence of endogenous right-hand-side variables and variables measured with error. However, this solution has poor finite sample properties on bias and precision when “the lagged levels of the series are only weakly correlated with subsequent first-differences, so that the instruments available for the first-differenced equations are weak” (Bond *et al.*, 2001). Blundell and Bond (1998) show that, in this case, Arellano and Bond’s (1991) solution has a large downward finite-sample bias. This problem occurs when the time series are persistent and the number of time series observations is small. An alternative solution would be to implement a system estimation, for first-differences and levels, as suggested by Blundell and Bond (1998). Bond *et al.* (2001) argue that this is the best solution to estimate growth regressions. In all our income regressions we used the GMM procedure proposed by Blundell and Bond (1998). The instrument set used is defined in the note to Table 5.

Table 5: Income Regressions, 10 year data

Variable	Edu			PEduFD		PEduAB	PEduBB
LagGDP	0.816** (0.073)	0.739** (0.060)	0.803** (0.074)	0.866** (0.081)	0.718** (0.087)	0.741** (0.091)	0.773** (0.088)
ΔS	0.133* (0.055)	0.221** (0.063)	0.133* (0.060)	0.086 (0.064)	0.192 [†] (0.116)	0.150 (0.113)	0.114 (0.105)
LagS	0.064* (0.029)	0.083** (0.030)	0.070* (0.031)	0.043 (0.037)	0.094* (0.039)	0.086* (0.042)	0.070 [†] (0.041)
KI	0.010** (0.003)	0.007* (0.004)	0.010** (0.003)	0.010** (0.003)	0.007 [†] (0.004)	0.007 [†] (0.004)	0.009* (0.004)
LagKI	0.004 (0.002)	0.003 (0.002)	0.004 (0.002)	0.004 [†] (0.003)	0.004 [†] (0.002)	0.004 [†] (0.002)	0.005* (0.002)
SF	-0.085 (0.083)	-0.086 (0.083)	-0.068 (0.077)	-0.093 (0.098)	-0.044 (0.061)	-0.054 (0.070)	-0.064 (0.078)
Before			-0.077 [†] (0.042)				
Last			-0.012 (0.016)				
LastC			0.003 (0.028)				
After			0.003 (0.016)				
Wald joint	1812.854**	1213.157**	2195.225**	1754.227**	1105.778**	1156.474**	1321.363**
Wald time	31.625**	29.209**	29.580**	39.358**	31.181**	29.224**	24.264**
Sargan	38.488	31.582	39.134	42.062	29.343	29.679	31.026
Sargan-df	36	28	32	36	28	28	28
DifSargan	19.257	17.760	19.580	23.159 [†]	17.928	17.084	14.491
AR(1)	-3.344**	-3.134**	-3.264**	-3.136**	-3.034**	-3.078**	-3.100**
AR(2)	1.146	1.070	1.035	1.161	0.942	1.101	1.223
Nobs	339	339	339	339	339	339	339
Ncountries	92	92	92	92	92	92	92

Significance levels : \dagger : 10% * : 5% ** : 1%. (Standard errors in parentheses)

The dependent variable is LogGDP. All regressions include time dummies. The instruments: (i) first-difference equations: level of LogGDP lagged two periods and earlier, levels of S, Ki and SF lagged one period and earlier till 5 lags; (ii) level equations: first-difference of LogGDP lagged one period, contemporaneous first-difference of S, Ki and SF. In columns 2, 5, 6, and 7 education is instrumented for measurement error. Alternative instruments: (i) S lagged three periods and earlier; (ii) first-difference of S lagged two periods.

Table 5 presents the results of the estimation of equation (9) using the data at 10 year intervals, while Table 10 shows the same estimations for the 5 year interval data. We estimate the model using different measures of education: *Edu*, *PEduFD*, *PEduAB* and *PEduBB*. The regressions are globally significant, and the time dummies included are jointly significant. The error term in equation (9) is not serially correlated, and the Sargan

tests in each regression do not reject the instruments used. The difference Sargan test also does not reject the estimation of the model by the system procedure, when compared with the first-differences procedure proposed by Arellano and Bond (1991).

As expected, the share of investment in GDP has a positive impact on economic growth, while State failure affects negatively growth. The coefficient on $LagGDP$ confirms the hypothesis of conditional convergence. While on the 10 year data it seems reasonable to assume that we do not have a unit root on income, we cannot claim the same for the 5 year data. We will concentrate our attention on the 10 year interval data, Table 5.

When we use the original variable, Edu , we conclude that both the level and the change in education are relevant for economic growth. The effect of education increases significantly, with the long term effect being 72%⁵, and the contemporaneous response of income being 13%. Following Bond *et al.* (2001), we instrument in column 2 for measurement error on education. For the equations in first-differences we use education lagged three periods and earlier, using at most 5 lags, while for the equations in levels we use first-differences of education lagged 2 periods. However, instrumenting for measurement error would be wrong if the measurement error in education is serially correlated. This is precisely our claim, which implies that these estimates are incorrect. In column 3 we include as regressors the census type of variables described above. The surprising result is that the regressor *Before* is significant after we control for education. Somehow this variable is capturing some feature of the country not included in the specific country effect. Apparently, those countries that take longer till the moment when they implement the first census or survey, have lower income growth. Countries' economic performance seems to be negatively correlated with the implementation of census.

In columns 4 till 7 we use the alternative measures of education described above. First, we estimate equation (9) using $PEduFD$ as the measure of education, columns 4 and 5. In column 5 we instrument for measurement error on education. Our assumption, and what distinguish column 5 from column 2, is that the measurement error in $PEduFD$ is not serially correlated. Our claim is that the transformation that we implemented to generate this variable filters out the systematic component of the measurement error in education. The same estimation is implemented using the variables $PEduAB$ and $PEduBB$, columns 6 and 7 respectively.

When we instrument $PEduFD$ for measurement error we conclude that both the level and the change in education are relevant for economic growth. The contemporaneous impact of an increase in education is around 19%, and the long-run effect is 33%. Using the variables $PEduAB$ and $PEduBB$ the results point in the same direction, but the coefficient on ΔS become statistically insignificant. The level effect stays economically and statistically significant.

5 Conclusion

Our analysis of Barro and Lee (2001) education data reveals that there is a systematic difference between the information collected from census or surveys, and the education

⁵The long-run multiplier is defined as $\frac{\beta_1 + \beta_2}{1 - \alpha}$.

data that results from the perpetual inventory method. On average, this method underestimates education by about one fifth of a year every five year period. This has an impact on the results for the growth regressions. Once we control for the source of information, and we take into account for measurement error in our variables, we conclude that both the level and the change in education are relevant for the growth process. However, alternative specifications and data intervals makes a difference for the size of the effects. Further research is need in order to make proper use of the knowledge on the systematic difference between census and non-census data.

Further research could also include the reestimation of the data on education. Using both the backward and the forward flow, the missing values can be reestimated using the average of both predictions, not only the weighted average between the linear interpolation and the forward prediction. However, the estimation of the missing values after the last census, and before the first census, would still be estimated the same way. It would also be important to estimate educational values taking into account for different survival rates according to the educational attainment. On this topic, Barro and Lee (2001) state that “the limitation of the data on age-specific education levels and mortality rates by age group do not allow us to compute specific mortality rates of population by levels of education.”

References

- Aghion, Philippe and Peter Howitt (1998), *Endogenous Growth Theory*, MIT Press, Cambridge, MA.
- Arellano, Manuel (2003), *Panel Data Econometrics*, Oxford University Press, New York.
- Arellano, Manuel and Stephen Bond (1991), ‘Some tests of specification for panel data: Monte carlo evidence and an application to employment equations’, *Review of Economic Studies* **58**, 277–297.
- Barro, Robert J. and Jong Wha Lee (1993), ‘International comparisons of educational attainment’, *Journal of Monetary Economics* **32**, 363–394.
- Barro, Robert J. and Jong Wha Lee (1996), ‘International measures of schooling years and schooling quality’, *American Economic Review* **86**, 218–223.
- Barro, Robert J. and Jong Wha Lee (2001), ‘International data on educational attainment: Updates and implications’, *Oxford Economic Papers* **3**, 541–563.
- Barro, Robert J. and Xavier Sala-I-Martin (1999), *Economic Growth*, MIT Press, Cambridge, Massachusetts.
- Benhabib, J. and M. Spiegel (1994), ‘The role of human capital in economic development: Evidence from aggregate cross-country data’, *Journal of Monetary Economics* **34**, 143–173.

- Blundell, Richard and Stephen Bond (1998), ‘Initial conditions and moment restrictions in dynamic panel data models’, *Journal of Econometrics* **87**, 115–143.
- Bond, Stephen R., Anke Hoeffler and Jonathan Temple (2001), ‘GMM estimation of empirical growth models’, Centre for Economic Policy Research Discussion Paper No. 3048.
- Bowsher, Clive G. (2002), ‘On testing overidentifying restrictions in dynamic panel data models’, *Economics Letters* **77**, 211–220.
- Cohen, Daniel and Marcelo Soto (2001), ‘Growth and human capital: Good data, good results’, OECD Development Centre Technical Paper No. 179.
- de la Fuente, Angel and Rafael Doménech (2002), ‘Human capital in growth regressions: How much difference does data quality make?’, Working paper, Instituto de Análisis Económico, Barcelona.
- Doornik, J.A. (2002), *Object-Oriented Matrix Programming Using Ox, 3rd Ed.*, Timberlake Consultants Press and Oxford, London. www.nuff.ox.ac.uk/Users/Doornik.
- Doornik, J.A., M. Arellano and S. Bond (2002), *Panel Data Estimation Using DPD for Ox*, Oxford, London. www.nuff.ox.ac.uk/Users/Doornik.
- Hanushek, Eric and Dennis Kimko (2000), ‘Schooling, labor force quality, and the growth of nations’, *American Economic Review* **90**(5), 1184–1208.
- Krueger, Alan and Mikael Lindahl (2001), ‘Education for growth: Why and for whom?’, *Journal of Economic Literature* **39**(4), 1101–1136.
- Kyriacou, G. (1991), ‘Level and growth effects of human capital: A cross-country study of the convergence hypothesis’.
- Lucas, Robert (1988), ‘On the mechanics of economic development’, *Journal of Monetary Economics* **22**(1), 3–42.
- Marshall, Monty G. and Keith Jagers (2000), ‘Polity IV project: Dataset users manual’, mimeo, Center for International Development and Conflict Management (CIDCM), University of Maryland.
- Nehru, V., E. Swanson and A. Dubey (1995), ‘A new database on human capital stocks in developing and industrial countries: Sources, methodology and results’, *Journal of Development Economics* **46**, 379–401.
- Nelson, Richard and Edmund Phelps (1966), ‘Investment in humans, technology diffusion and economic growth’, *American Economic Review* **56**(2), 69–75.
- Summers, Robert and Alan Heston (1991), ‘The penn world table (mark 5): An expanded set of international comparisons, 1950-1988’, *Quarterly Journal of Economics* **106**(2), 280–315. (the new dataset, mark 6.1, is available at <http://pwt.econ.upenn.edu/>).

Teulings, Coen and Thijs Van Rens (2003), ‘Education, growth, and income inequality’, Tinbergen Institute Discussion Paper 02-001/3.

Windmeijer, Frank (2000), ‘A finite sample correction for the variance of linear two-step GMM estimators’. IFS working paper W00/19 (www.ifs.org), London: The Institute for Fiscal Studies.

Woessmann, A. (2000), ‘Specifying human capital: A review, some extensions, and development effects’, Kiel Institute of World Economics Working Paper No. 1007.

Table 6: Description of Census variables

Variable	Mean	% of zeros	Min.	Max.
Census	0.32	0.68	0	1
Before	0.23	0.88	0	5
Last	0.64	0.62	0	6
LastC	0.38	0.80	0	6
After	0.95	0.62	0	8
N		985		

Table 7: Distribution of Census per Country

Number of census	Countries	Per cent
1	25	22
2	28	24
3	33	28
4	22	19
5	6	5
7	1	1
8	1	1
Total	116	100

Table 8: Description of Education

Sample	Mean	Std. Dev.	Min.	Max.	N
Full sample	5.03	2.88	0.09	12.05	985
Year 1960	3.78	2.59	0.12	9.73	104
Year 1965	3.9	2.56	0.17	9.74	104
Year 1970	4.28	2.7	0.2	10.24	106

Continued on next page...

... table 8 continued

Variable	Mean	Std. Dev.	Min.	Max.	N
Year 1975	4.52	2.75	0.09	11.27	110
Year 1980	4.99	2.86	0.26	11.86	111
Year 1985	5.31	2.8	0.49	11.57	112
Year 1990	5.84	2.84	0.65	11.74	116
Year 1995	6.07	2.8	0.76	11.89	111
Year 2000	6.33	2.82	0.84	12.05	111

Table 9: Description of Education for Census observations

Sample	Mean	Std. Dev.	Min.	Max.	N
Full sample	5.2	2.9	0.09	11.89	313
Year 1960	3.67	2.06	0.12	7.85	51
Year 1965	4.07	2.34	1.04	9.57	21
Year 1970	4.88	2.64	0.2	10.24	57
Year 1975	4.3	2.99	0.09	11.27	39
Year 1980	5.56	2.94	0.26	11.86	67
Year 1985	5.57	2.68	1.34	10.76	22
Year 1990	7.11	2.68	1.55	11.74	45
Year 1995	8.60	2.29	5.12	11.89	11

Table 10: Income Regressions, 5 year data

Variable	Edu			PEduFD		PEduAB	PEduBB
LagGDP	0.926** (0.047)	0.926** (0.049)	0.927** (0.045)	0.946** (0.044)	0.921** (0.051)	0.924** (0.048)	0.934** (0.048)
ΔS	0.043* (0.022)	0.074 (0.057)	0.037 (0.024)	0.034 (0.026)	0.068 (0.053)	0.069 (0.073)	0.065 (0.075)
LagS	0.038* (0.016)	0.037 [†] (0.023)	0.039* (0.016)	0.027 [†] (0.016)	0.041 [†] (0.024)	0.040 [†] (0.024)	0.037 (0.024)
KI	0.005* (0.003)	0.006* (0.003)	0.005* (0.003)	0.006* (0.003)	0.005 [†] (0.003)	0.005 [†] (0.003)	0.006 [†] (0.003)
LagKI	0.002 (0.003)	0.002 (0.003)	0.002 (0.003)	0.002 (0.003)	0.002 (0.003)	0.002 (0.003)	0.002 (0.003)
SF	-0.096 (0.082)	-0.093 (0.081)	-0.115 (0.098)	-0.118 (0.089)	-0.125 (0.081)	-0.118 (0.081)	-0.104 (0.087)
Before			-0.024 (0.020)				
Last			0.014 (0.009)				
LastC			-0.010 (0.011)				
After			0.007 (0.009)				
Wald joint	1872.535**	3026.874**	1912.394**	2223.998**	2267.356**	2596.071**	2873.966**
Wald time	48.475**	56.075**	30.775**	43.420**	49.900**	48.939**	38.024**
Sargan	92.074	87.462	84.062	89.419	86.270	86.236	87.655
Sargan-df	112	96	108	112	96	96	96
DifSargan	3.947	14.716	1.294	0.674	17.816	18.771	21.318
AR(1)	-3.054**	-2.966**	-3.042**	-3.048**	-3.041**	-2.940**	-2.879**
AR(2)	0.995	1.019	1.002	0.987	0.972	0.917	0.893
Nobs	694	694	694	694	694	694	694
Ncountries	95	95	95	95	95	95	95

Significance levels : † : 10% * : 5% ** : 1%. (Standard errors in parentheses)

The dependent variable is LogGDP. All regressions include time dummies.

For the definition of the instruments see the note to Table 5.