CROSS-COUNTRY DATA ON THE QUANTITY OF SCHOOLING: A SELECTIVE SURVEY AND SOME QUALITY MEASURES

Angel de la Fuente[*] *and Rafael Doménech* [**]

D-2013-06

September 2013

[*]Instituto de Análisis Económico (CSIC) and Barcelona GSE
[**] BBVA Research and Universidad de Valencia

This document is available at: http://www.sepg.pap.minhap.gob.es/sitios/sepg/es-ES/Presupuestos/Documentacion/Paginas/Documentacion.aspx

The Working Papers of the Dirección General de Presupuestos are not official statements of the Ministry of Finance and Public Administrations.

Abstract

We survey a number of papers that have focused on the construction of cross-country data sets on average years of schooling. We discuss the construction of the different series, compare their profiles and construct indicators of their information content. The discussion focuses on a sample of OECD countries but we also provide some results for a large non-OECD sample.

Key words: human capital, growth, measurement error
JEL Classification: O40, I20, O30, C19

1. Introduction

This paper is part of a selective and critical survey of the recent literature on the measurement of the quantity and quality of human capital. It describes and compares a number of recent cross-country data sets on average years of schooling, with particular emphasis on an OECD sample, and constructs statistical measures of their information content. A companion piece will deal with efforts to measure educational quality, mostly by relying on the scores of standardized international tests.

The construction of homogeneous schooling series for broad samples of countries has been the main goal of a significant and growing number of papers over the last three decades. One of the earliest attempts in this direction is due to Psacharopoulos and Arriagada (P&A, 1986) who, drawing on earlier work by Kaneko (1986), report data on the educational composition of the labor force in 99 countries and provide estimates of average years of schooling. In most cases, however, P&A provide only one observation per country. More recently, there have been many attempts to construct more complete data sets on educational attainment that provide broader temporal coverage and can therefore be used in growth accounting and other empirical exercises. These series are generally constructed using data from international compilations of attainment and/or enrollment data from UNESCO and other organizations and employing different procedures to build up stock estimates from enrollment data and/or to fill in missing stock observations. The relevant literature includes papers by Kyriacou (1991), Lau, Jamison and Louat (1991), Lau, Bhalla and Louat (1991), Nehru, Swanson and Dubey (NSD 1995), Barro and Lee (1993, 1996, 2001 and 2013), Cohen and Soto (2007), de la Fuente and Doménech (2002, 2006 and 2012), Lutz et al (2007) and Samir et al (2010).

Most of this literature has been surveyed in some detail in de la Fuente and Doménech (2006), where we also construct statistical measures of the information content of most of the data sets that were available at the time. The present paper updates and extends our earlier work, focusing on four data sets that appear ex-ante to be potentially most useful for empirical researchers because of their quality and coverage. We focus in particular on the most recent available versions of the schooling series constructed by Barro and Lee (B&L), Cohen and Soto (C&S), Lutz, Samir et al (L&S+) and de la Fuente and Doménech (D&D), working mostly with an OECD sample.[1] Section 2 briefly reviews the methodology used to construct these series, with further details in Appendix 1. The different data sets are compared with each other in section 3 and measures of their information content are constructed

---

[1] For Barro and Lee, we use version 1.2 (released in 2011) of the data set described in B&L (**2013**), which is available at http://www.barrolee.com; for D&D we use version 3.0, as described in D&D (2012), which can be downloaded from http://ideas.repec.org/p/aub/autbar/911.12.html; for L&S+, we work with an unpublished "current working version" supplied in 2012 by K. C. Samir, to whom we are grateful, and for C&S we use an updated version of their (2007) data set which was downloaded from http://soto.iae-csic.org/Data.htm in 2012. Since the C&S data come only at 10-year intervals, we use linear interpolation to complete the quinquennial series with which we work. We thank K. C. Samir and M. Soto for providing the latest available versions of their data.

in section 4. Section 5 concludes and Appendix 2 extends some of the work to a large sample of non-OECD countries.

2. The construction of some schooling data sets

Barro and Lee (B&L, 1993) construct attainment series for a large number of countries covering the period 1960-85 by combining data on enrollment rates with census information, both taken primarily from UNESCO compilations. To estimate attainment levels in years for which census data are not available, they use a short-cut perpetual inventory procedure that can be used to estimate changes from nearby (either forward or backward) benchmark observations using data on enrollments and the age distribution of the population. This data set, which has been extensively used in the empirical growth literature, has been revised, updated and extended in a series of papers by the same authors (B&L, 1996, 2001 and 2013). It has also been criticized by other researchers, who have constructed alternative schooling series that attempt to improve the signal-to-noise ratio in the data.

Barro and Lee's work has focused on expanding data coverage, improving the procedure used to fill in gaps in the census data and providing an increasingly detailed breakdown of the information by sex and age group while continuing to rely on Unesco compilations as their main source of raw attainment and enrollment data. Other authors, however, have relied increasingly on other sources in an attempt to eliminate anomalies in the data arising in all likelihood from changes in classification criteria that are hard to detect in the supposedly homogenized Unesco data. After documenting the problems found in the most widely used schooling series, de la Fuente and Doménech (D&D 2002, 2006 and 2012) construct new attainment data for a sample of 21 OECD countries. Mistrustful of the homogeneity of UNESCO's compilations, these authors rely primarily on OECD and national sources and focus on constructing plausible time profiles for attainment in each country. Cohen and Soto (2002, 2007) refine B&L's fill-in procedure by making full use of the available census data on attainment by age group in order to allow survival rates to differ across age groups (see below). They also incorporate new survey data from the OECD's in house database and attempt to mitigate the problem caused by changes in classification criteria by disregarding census observations that may be affected by such changes and relying instead on backward projections based on more recent census information.[2] This approach is taken to the extreme by Lutz et al (2007), who construct schooling series going back to 1970 by projecting backward data taken exclusively from a single census, that of 2000. Samir et al (2010) revise these series and extend the Lutz et al (2007) data set forward using the same basic data for around 2000 and a similar methodology to project attainment forward. As noted, we will work with an unpublished current working version of these series to which we will refer as the L&S+ data set (for Lutz, Samir et al).

In this section we will take a closer look at the content of these four series and at the methodology used to construct them. All of these data sets provide information on the fraction of the adult population (understood as those of age 15 or 25 and over) that has attained each of several possible

---

[2] The authors are not very explicit on how classification changes are detected. They seem to rely on changes in the duration of the different school cycles as reported in Unesco's Yearbook (see p. 53).

educational levels and on the average years of schooling of such population, in some cases disaggregated by sex and/or by age group.

The seven levels of schooling considered by B&L and C&S are: no schooling and complete and incomplete primary, secondary and higher education.[3] D&D consider two cycles (lower and upper) of secondary and higher education and L&S+ distinguish only between no schooling and primary, secondary and higher education. They include persons with incomplete lower secondary training in their primary category and those with incomplete short college careers are counted as having only secondary attainment.

In most cases, average years of schooling are calculated using attainment shares and the theoretical durations of the different school cycles in each country. B&L (1993 and 1996) use constant durations, taken from UNESCO's Statistical Yearbook and in principle applying to 1965.[4] In their more recent work, the same authors (B&L, 2000 and 2013) allow for changes in durations over time and take into account that such changes are incorporated only gradually into the stock of human capital as the affected cohorts enter the adult population. D&D and C&S apply recent theoretical durations to the entire period. D&D (2012) take their standard durations from national sources,[5] while C&S (2007) seem to rely on Unesco data (see footnote 1 in p. 53). L&S+ also rely on Unesco data on durations for 2000 or a nearby year but they use a slightly different approach. Instead of using standard durations directly, they rely on these data to estimate (in an admittedly ad-hoc manner) the average years spent in school by persons included in each attainment category.[6]

As noted, Barro and Lee rely primarily on UNESCO and other UN compilations of census/survey data but also take some information from the web pages of Eurostat and several national statistical institutes. On the other hand, they disregard OECD data on educational attainment claiming that they may not be compatible with other sources because they are generally based on (labor force or other) surveys rather than on full censuses as most of the UNESCO data. As a result, they argue, OECD data tends to exclude people of retirement age and thus refers to a different population group than the census data (typically 25-64 in the OECD vs. 15+ or 25+ in most censuses). They also note that these data are based on a different classification scheme that, among other things, lumps together all persons with less than upper secondary attainment (B&L 2001, pp. 558-60). Cohen and Soto (2007) by

---

[3] Barro and Lee include in the "incomplete secondary" category those who have started the first cycle of secondary education but not progressed beyond this level, and in the "complete secondary" those who have started but not necessarily completed upper secondary schooling (and have not started post-secondary education). These authors include "short" college-level diplomas in incomplete higher education, together with incomplete longer degrees. In the case of Cohen and Soto, it is not clear whether they follow the same convention or define their complete and incomplete secondary and higher education categories with a different criterion.

[4] When no data are available on the separate durations of the two cycles of secondary, Barro and Lee assign half the total duration of secondary to each cycle. Incomplete primary also gets assigned ½ of the duration of complete primary education. For incomplete and complete higher education they use 2 and 4 years in all countries. C&S (20007) assign half of a level's theoretical duration to the incomplete category.

[5] In the case of Spain, we take into account changes in durations as a result of educational reforms.

[6] This sort of correction seems particularly pertinent in the case of L&S because they use rather broad educational attainment categories. As the authors note, a person included in L1 in Mexico could have stayed in school anywhere between 1 day and 9 years minus one day. Since attributing 9 years of schooling to all these people would surely overestimate their attainment, it seems preferable to take an intermediate figure for the average schooling of the population with primary attainment even if this cannot be based on precise data. See Samir et al, pp. 403-4.

contrast, rely primarily on OECD data for those countries for which they are available. For most OECD member countries, their estimates are based only on OECD data for the nineties, ignoring a large amount of information available in other sources. L&S+ (2007 and 2010) only use census or survey data for the year 2000, broken down by age group and taken from censuses (mostly as compiled by UNESCO) or labor force or demographic and health surveys. This has the advantage of ensuring that a consistent attainment classification is applied (retroactively) to all cohorts throughout the sample period, but may bias the results in countries with significant migration flows. Finally, D&D (2012) rely primarily on national sources and make only occasional use of UNESCO data and other compilations.

In both B&L and C&S, there is some ad-hoc filtering of original census data. Cohen and Soto disregard earlier censuses when they suspect there have been changes over time in classification criteria and proceed by projecting backward more recent and presumably more homogeneous census data. As noted, Barro and Lee disregard OECD data and in their 2013 paper they adjust some census observations that seem to be "off trend" (in the cases of Canada 1975, France 1955 and 1990, Italy 1980 and Korea 1990, see B&L 2012).

Table 1: Key features of several schooling data sets

	B&L (1993)	B&L (1996)	B&L (2000)	C&S (2007)	B&L (2012)	L&S+ (2007/10)	D&D (2012)
period	1960-85	1960-90	1960-2000	1960-2010	1950-2010	1970-2000/10	1960-2010
frequency	5 yrs.	5 yrs.	5 yrs.	10 yrs.	5 yrs.	5 yrs.	5 yrs.
population group	25+	15+, 25+	15+, 25+	15+, 25+	15+, 25+	15+, 25+	25+
disaggregation by	sex	sex	sex	-	sex&age	sex&age	-
# of countries							
with complete data	106	105	109	95	146	120	21
with incomplete data	23	21	33		45		
% of direct observations	40.2%	35.1%	27.7%	24.4%	25.0%	14.3%	58.0%
basic fill-in procedure	Perpetual inventory	Perpetual inventory	Perpetual inventory	Projections with detailed data by age group	Projections with detailed data by age group	Projections with detailed data by age group	Linear interpolation+ backward projections w/ detailed data by age group
enrollment variable used in fill-in procedure	gross enrollment ratio	net enrollment ratio	gross enrollment ratio adjusted for repeaters	estimated net intake ratios	gross enrollment ratio adjusted for repeaters	none	none
survival probs. vary with:							
educational level	no	no	no	no	partially	yes	partially
age	no	no	no	yes	yes	yes	yes
allow for changes in durations	no	no	yes	no	yes	no	only for Spain

Table 1 shows the geographical and time coverage of the relevant studies and summarizes some of their key features. All studies begin by collecting "direct" census or survey data, which make up between 14% and 58% of the potential observations. Missing observations are then estimated using either interpolation or some sort of fill-in procedure to construct forward or backward projections using nearby census observations and possibly enrollment data. In the first three versions of the Barro and Lee data set, this is done using a short-cut perpetual inventory procedure in which the attainment of the adult population at time *t* is estimated as a weighted average of the attainment of the same age group in a nearby census year and the attainment of new entrants into the desired age group during the intervening period, which is estimated using enrollment data, possibly adjusted for repeaters and dropouts, or net intake rates (the fraction of the relevant population that enters each educational cycle).[7]

Cohen and Soto (2007) improve on this procedure by using the available detail on attainment by age group to construct more accurate forward and backward projections. The main advantage of this procedure is that it implicitly allows survival probabilities (in the period elapsed since the census observation that is used as a starting point) to vary across cohorts, whereas the short-cut perpetual inventory procedure used in previous papers imposed a common survival rate for the entire adult population. In the latest version of their data set, Barro and Lee (2013) adopt this methodology and introduce a further refinement that allows survival probabilities to vary with the level of education for the oldest cohorts. D&D (2012) employ this refined procedure in the backward projections they use to extend the series in those countries for which there are no data in the earlier years of the sample period, but rely on linear interpolation to fill in gaps between available census data.

L&S (2007 and 2010) use an extrapolation procedure similar to the one used by C&S (2007) and B&L (2013). L&S+ deviate from the standard practice in other studies in that they rely on a single census, which is projected backward and forward.[8] While having some obvious drawbacks, this procedure does have the advantage of avoiding problems arising from changes in classification criteria over time. Some of the details of the projection procedure also differ from previous studies. In principle, L&S+ allow mortality rates to differ across age and attainment groups and over time. Unlike B&L and C&S, they do not use enrollment data when estimating the attainment level of the youngest and oldest cohorts at each point in time. Instead, they basically extrapolate the cross-cohort attainment pattern found in their basic data to estimate the attainment of new entrants into the adult population and that of unobserved age groups that are part of the oldest, open-ended population segment (typically the group 65+).

In most cases, the basic fill-in procedure is applied using a coarse classification into four broad educational levels (no schooling, some primary, some secondary and some higher education) and the

---

[7] For the details of this procedure and the refinements introduced by different authors, see Appendix 1.

[8] The forward projections are constructed under several scenarios that incorporate different assumptions on fertility and migration rates and on the evolution of educational attainment in younger cohorts. The data we use for 2010 seem to be based on the central scenario ("global education trend") and, at any rate, they will not be very sensitive to such assumptions since the relevant birth rates are known as of 2000 and the rest of the assumptions affect only the very youngest 15+ cohorts in 2010.

finer breakdown is completed ex-post using estimates of completion ratios, i.e. of the fraction of each population subgroup that has actually completed each school cycle.

3. A closer look at the data

In this section we compare our attainment series (D&D, 2012) with the latest available versions of the C&S, B&L and L&S+ data sets, restricted to our sample of 21 OECD countries.[9] We find that there are significant differences across these four sources in terms of both their cross-section and their time series profiles. Another cause for concern is that some series display extremely large changes in attainment levels over periods as short as five years (particularly at the secondary and tertiary levels).

Table 2: Correlation among alternative estimates of average years of schooling over common observations in the OECD21 sample, quinquennial data in levels/growth rates

|	C&S	L&S	D&D
Barro and Lee (B&L 13)	0.819/0.336	0.732/0.569	0.801/0.315
Cohen and Soto (2007)		0.839/0.702	0.928/0.581
Lutz, Samir et al ()			0.904/0.847

Table 2 shows that the overall correlation (computed over common observations) between different estimates of average years of schooling is reasonably high when the data are measured in levels and considerable lower when we work with growth rates. The high overall correlation across the series in levels, moreover, hides significant discrepancies across them. As an example, Figure 1 compares B&L's (2013) estimates of years of schooling in 2000 with our own (D&D, 2012), after normalizing each series by the corresponding sample average. As can be seen in the figure, the discrepancies between the two sources are very large for a number of countries. B&L provide much more optimistic estimates of relative attainment than we do in the cases of New Zealand (with a difference of 22.6 points between the two estimates in favor of B&L), Spain (+16.4) and Ireland (+15.6), and are much more pessimistic for Finland (-20.7), the UK (-13.5) and Austria (-12.1), to mention only the more extreme cases. These discrepancies substantially change the relative position of some countries within the attainment distribution. New Zealand, for instance, drops from the 2nd position to the 14th as we go from B&L to D&D, while Ireland goes from 5th to 16th and Finland rises from 20th to 11th.

Looking in greater detail at the different attainment series for a given country, the differences can also be quite significant. As an illustration, Figure 2 compares the four series of average years of schooling in the cases of Germany and Finland. For Germany, the C&S and D&D series roughly agree on their average levels and on the existence of a soft upward trend, while L&S+ paint a much flatter time profile at a significantly higher attainment level. Finally, B&L's series displays a completely different profile for the same country: after starting from a much lower level, these authors' estimate of German attainment rises rapidly during the second half of the sample period and converges to D&D's series in

---

[9] For a comparison between the three other data sets outside the OECD, see Appendix 2.

its final decade.[10] For Finland, the pattern is similar. C&S and D&D roughly agree, L&S+ is significantly more optimistic and B&L's series displays an implausible time profile, with surprising fluctuations in average years of schooling during the second half of the sample period.

Figure 1: Average years of schooling in 2000: B&L (2013) vs D&D (2012)
**Normalized years of schooling in 2000, B&L vs. D&D**
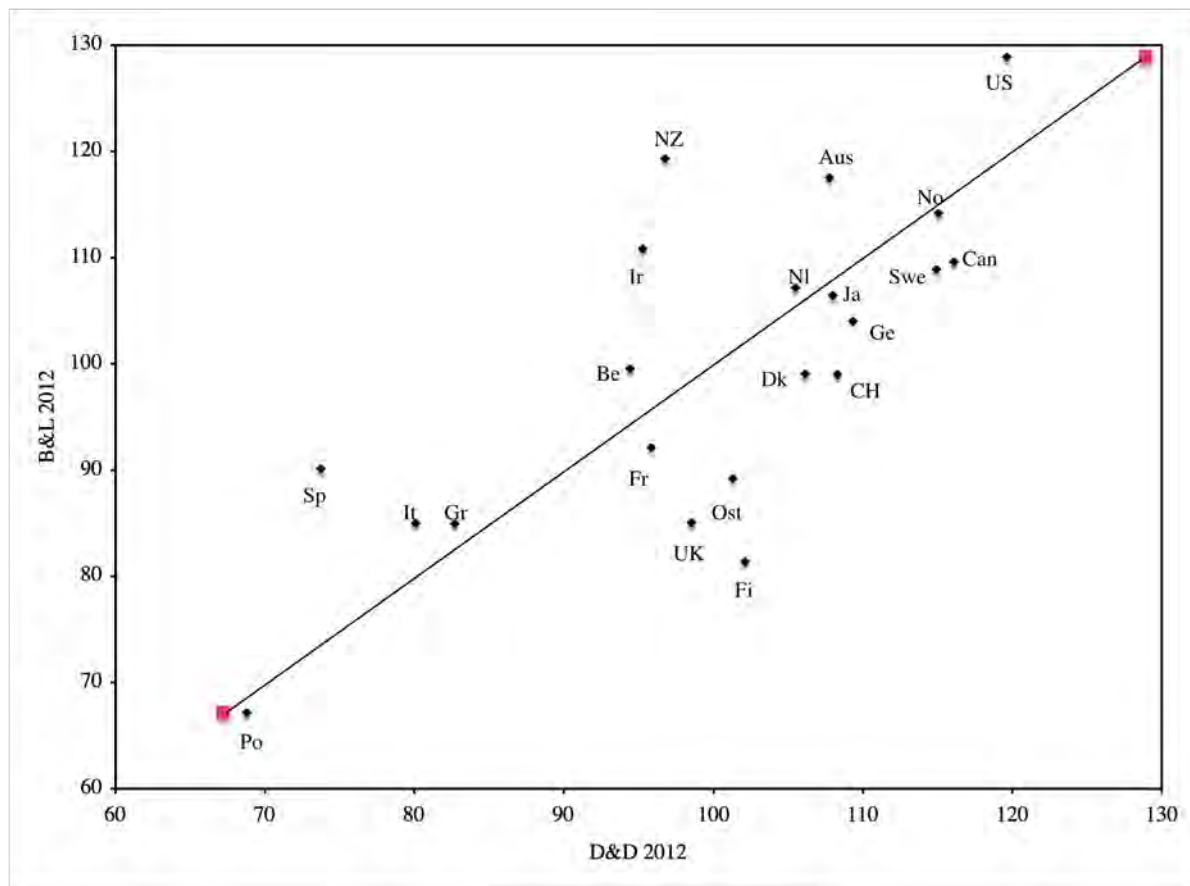
Legend: *Pr* = Portugal; *Sp* = Spain; *It* = Italy; *Gr* = Greece; *Be* = Belgium; *Ir* = Ireland; *Fr* = France; *NZ* = New Zealand; *UK* = United Kingdom; *Ost* = Austria; *Fi* = Finland; *Dk* = Denmark; *Nl* = Netherlands; *Aus* = Australia; *CH* = Switzerland; *Ja* = Japan; *Ge* = Germany; *Swe* = Sweden; *No* = Norway; *Can* = Canada; *US* = United States.

To compare the cross-section profiles of the different series of years of schooling in a somewhat more systematic manner, we begin by normalizing each of them by its contemporaneous sample mean and by calculating the average of these normalized figures during the period in which all four series overlap (1970-2010), which is shown in Table 3. Working with this summary indicator of average relative schooling over the entire sample period, Figure 3 shows the differences across sources, taking as a reference Barro and Lee's (2013) estimates. Figure 4 is constructed in the same way but working now with the observed variation in normalized schooling between 1970 and 2010.

---

[10] Our data for this country refer to West Germany until 1985 and to the united country thereafter. The same seems to be true for B&L (see their Appendix notes on Germany). On the other hand, C&S always refer to the entire country (see footnote 6 in p. 56) and the same must be true for L&S+ by construction since they work with the 2000 census. Our estimates suggest that attainment differences between East and West Germany at the time of unification were very small, so differences across data sets on the treatment of Germany should not make a big difference.

Figure 2: Average years of schooling according to different sources

a. Germany
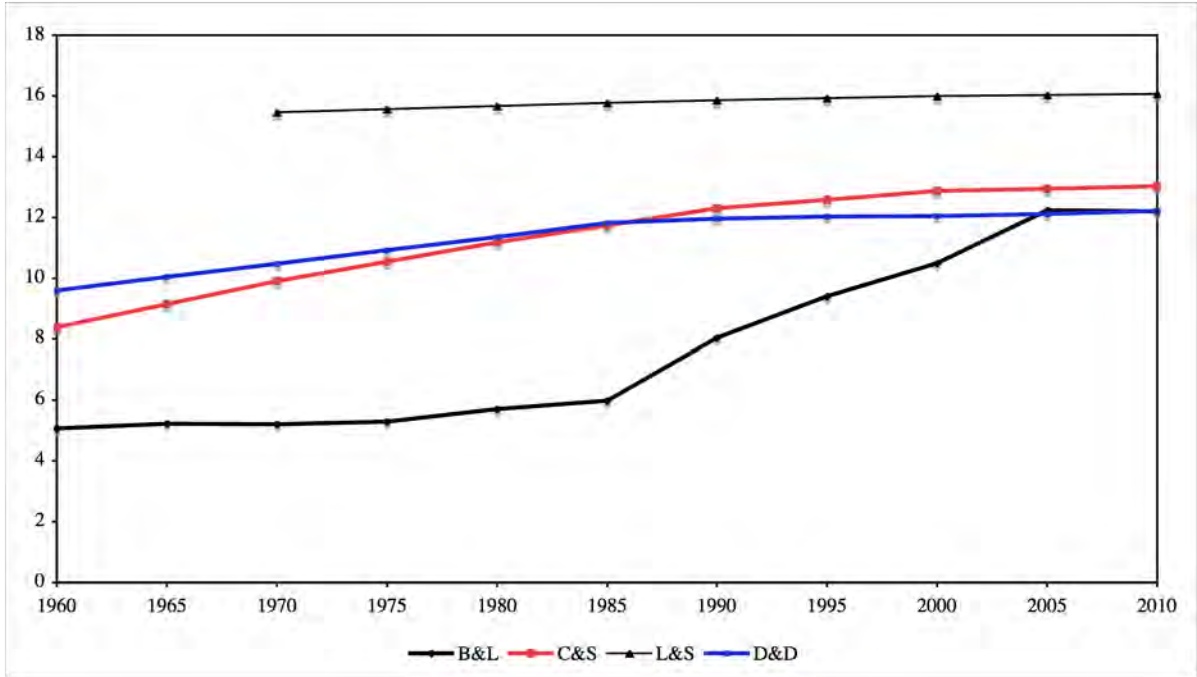
b. Finland
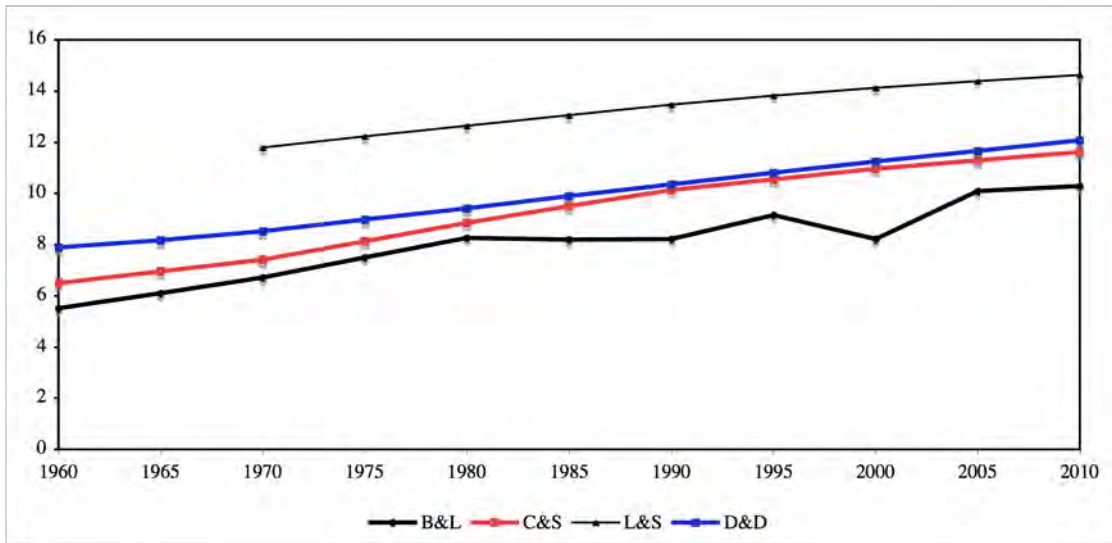
As in the case of Figure 1, some of the disagreements across sources are very important. For instance, Barro and Lee place Germany in the lower half of the distribution of attainment, with an average relative schooling index of 88 over the period 1970-2010, while all other sources place it in the upper tail of the distribution, with an index of around 120 or higher. The opposite happens in the cases of New Zealand and Ireland, where B&L's figures are much more optimistic than the rest.

Table 3: Normalized years of schooling
average value over the period 1970-2010

	B&L13	C&S	L&S	D&D12
USA	135.1	119.6	112.6	124.0
New Zealand	130.0	107.9	111.0	96.6
Australia	126.3	119.3	100.5	112.1
Norway	114.8	110.4	127.7	121.0
Canada	113.5	115.1	122.3	119.5
Ireland	112.0	88.2	90.5	93.3
Netherlands	108.6	101.5	100.6	105.0
Sweden	108.4	105.5	99.8	113.3
Switzerland	106.5	122.4	118.2	114.0
Japan	106.1	110.4	127.3	109.4
Denmark	104.4	107.5	124.8	111.9
Belgium	98.3	91.7	89.6	91.9
Finland	92.9	97.4	116.7	101.2
UK	88.4	111.9	84.9	94.3
Germany	88.0	118.6	139.0	115.0
Austria	87.5	99.8	118.9	105.3
Greece	85.5	78.0	69.6	77.9
France	80.8	89.1	77.9	91.4
Italy	78.9	78.9	66.8	74.3
Spain	71.8	74.3	52.1	67.5
Portugal	62.2	52.3	49.2	61.1
Average	100	100	100	100

- *Note:* Average of quinquennial observations. For C&S we interpolate between decennial observations to complete the quinquennial series prior to calculating the average.

Figure 3: Normalized years of schooling, differences with B&L (2013)
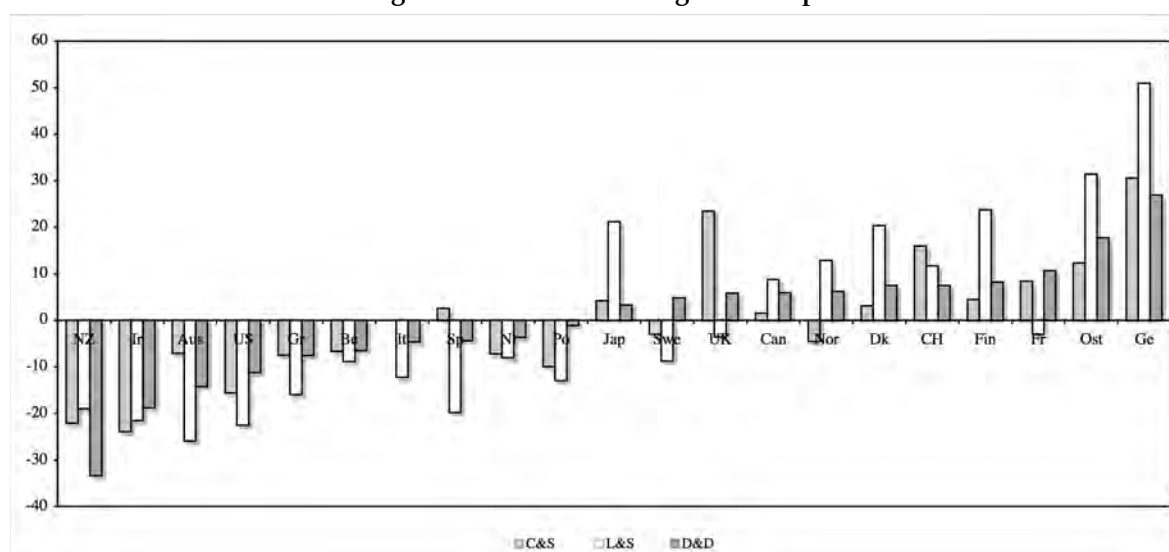**based on average normalized schooling over the period 1970-2010**

Figure 4: Variation in normalized years of schooling between 1970 and 2010, differences with B&L (2013)
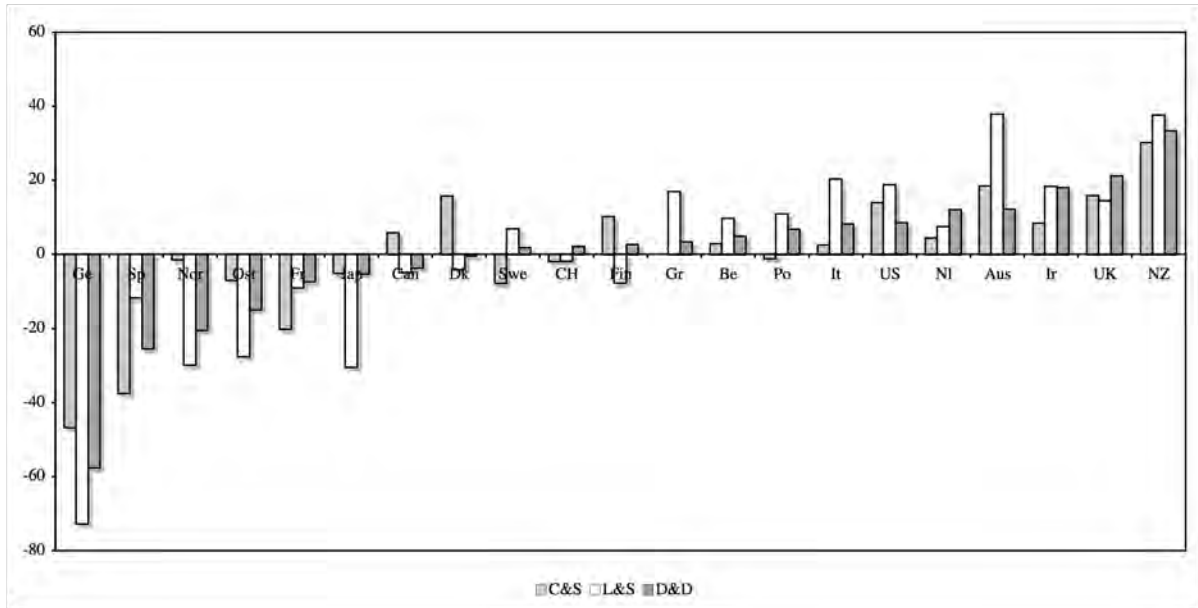

Table 2 suggests that the D&D, C&S and L&S+ series are somewhat closer to each other than to the B&L data set, which stands apart, displaying generally lower correlations with the other three sources than these have among themselves. Figures 3 and 4 tend to confirm this conclusion: there is broad agreement across the other three sources regarding at least the sign of the difference with the B&L series and quite often its magnitude, both in levels and in long differences (between 1970 and 2010). There are many exceptions to this pattern, however. For instance, L&S+ are considerably more optimistic about Japan, Denmark and Finland than the other three sources, which are relatively close to each other for these countries.

To construct a rough measure of the degree of agreement across series in levels, we will say that two sources agree for a given country if the maximum difference between them in terms of average normalized years of schooling is less than 5% of their average value. We find that there is no country for which all four sources agree. The highest degree of agreement (10 countries out of 21) is attained by comparing our data with the C&S series, and the lowest (2 countries) corresponds to the combination of B&L with L&S+. Table 4 shows the degree of pairwise agreement of the different series, measured by the percentage of cases in which the stated agreement criterion is satisfied.

Table 4: Degree of agreement between different pairs of normalized schooling series in levels

	C&S	L&S	D&D
Barro and Lee (B&L 13)	38%	10%	19%
Cohen and Soto (C&S)		24%	48%
Lutz, Samir et al (L&S)			24%

**Figure 5: Fitted distribution of the growth rate of years of schooling, different data sets
OECD21 sample**
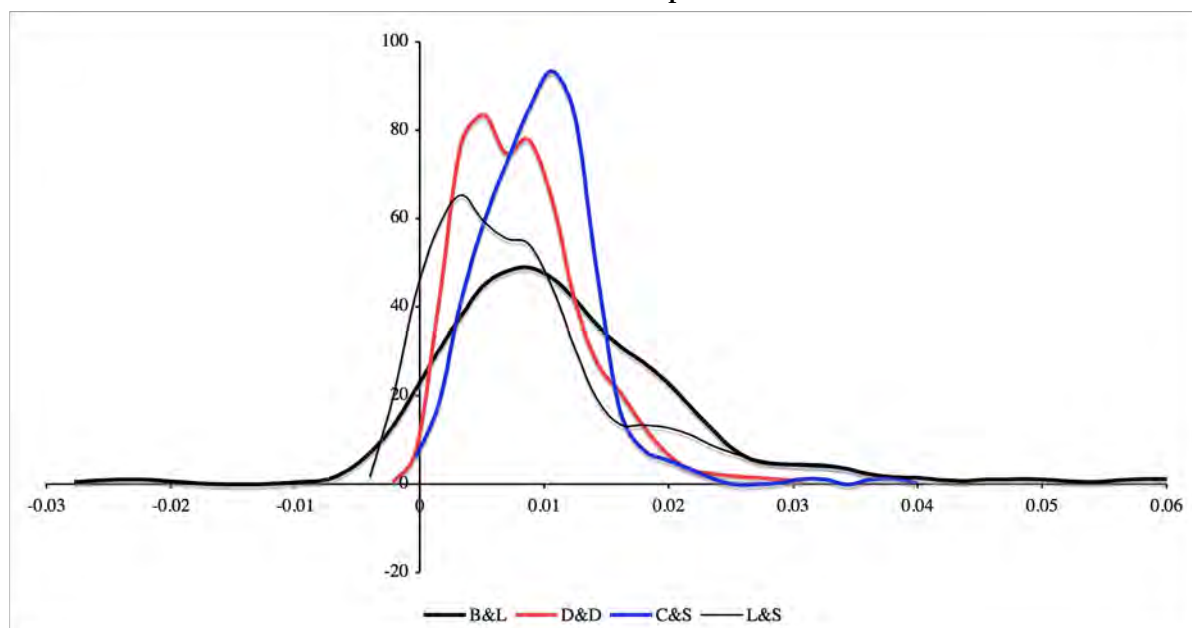
When we turn to the time profiles of the different data sets, C&S, D&D and L&S+ display a considerably smoother pattern than B&L. This is clearly illustrated in Figure 5, where we have plotted the fitted distribution of the annualized quinquennial growth rate of average years of schooling (using in each case all the available observations for the same OECD sample). The differences in the range of this variable across data sets are enormous: while our annual growth rates range between 0.04% and 2.90%, Barro and Lee's go from -2.52% to 5.92%; moreover, 6.2% of the observations in this last data set are negative, and 15.2% of them exceed 2%. As shown in Table 5, C&S and L&S+ occupy intermediate positions in terms of their range. The L&S+ series are very smooth by construction (see the country profiles in Appendix 3), but this is consistent with a fairly thick upper tail that comes largely from high growth rates of attainment in the Mediterranean countries during the early part of the period. On the other hand, there are several countries where L&S+ paint a very flat attainment profile that stands in contrast with other sources. The countries were this pattern is most clearly apparent are Japan and Norway.

Table 5: Range of different estimates of the growth rates of years of schooling

	D&D	C&S	L&S	B&L
max	2.90%	3.73%	4.06%	5.92%
min	0.04%	0.08%	-0.06%	-2.52%
% of negative observations	0.00%	0.00%	1.19%	6.19%
% of observations above 2%	1.90%	1.90%	10.71%	15.24%

As shown in Figure 6, an implausibly broad range of values (for the data in growth rates) is a common feature of all versions of the Barro and Lee 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.

Figure 6: Fitted distribution of the growth rate of years of schooling, different versions of the Barro and Lee data set
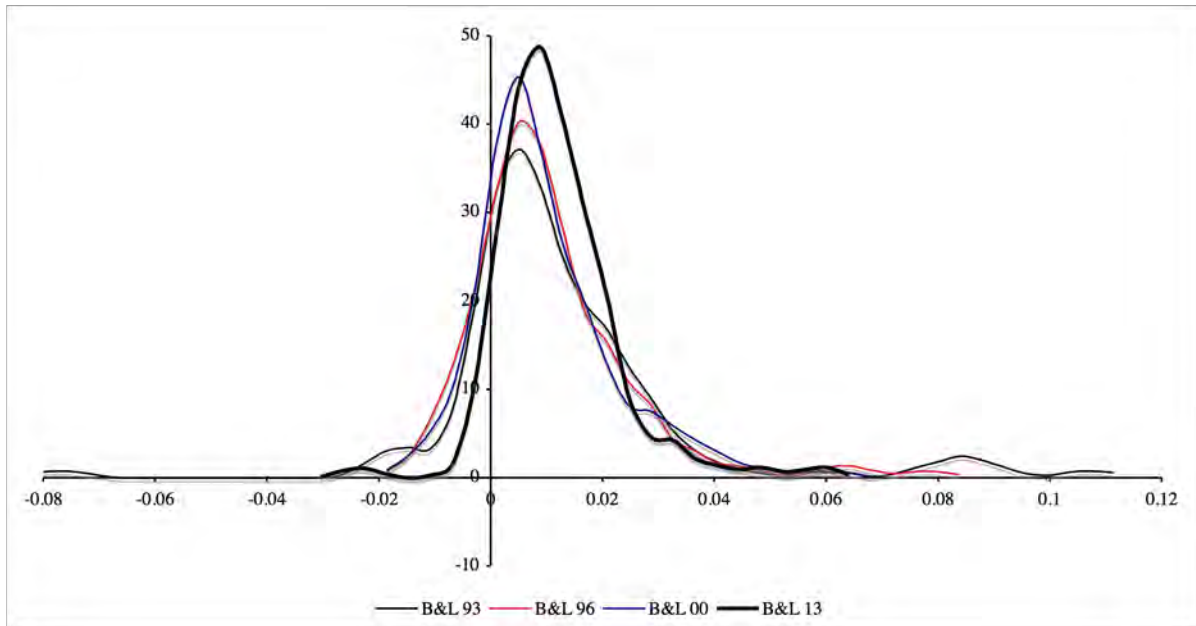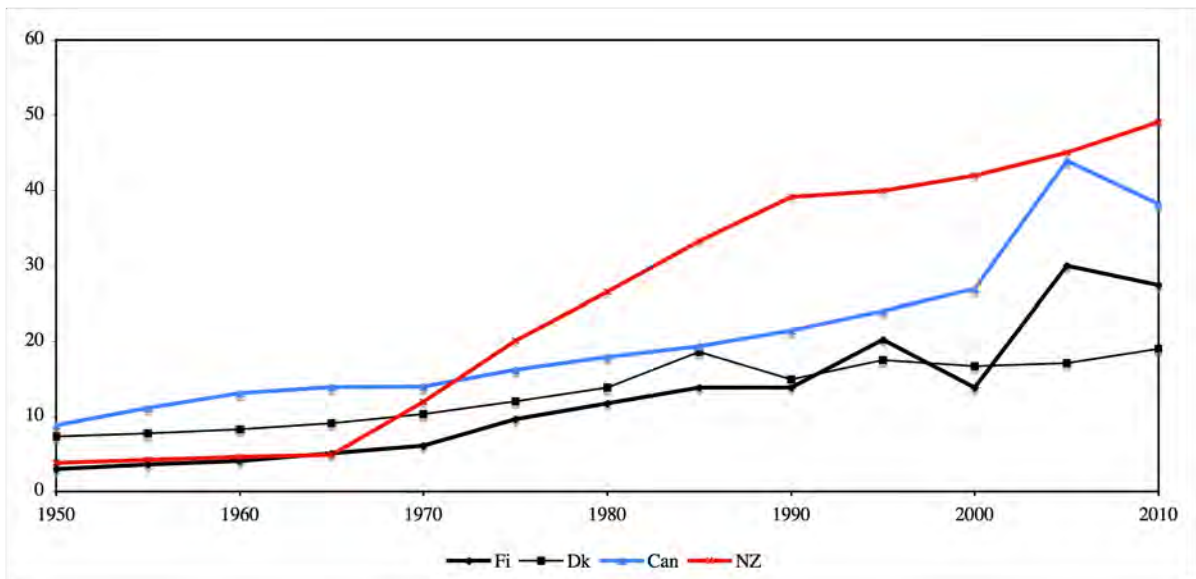

Figure 7: Evolution of university attainment levels in selected countries according to B&L (2013), % of the 25+ population

The volatility of the B&L series is a warning signal that it contains sharp breaks and implausible changes in attainment levels over very short periods. While this problem has become less severe with successive revisions of the data set, it remains even in its 2012 version. As an illustration, Figures 7 and 8 show the evolution of Barro and Lee's (2013) upper secondary and university attainment rates for the population over 25 in a number of countries that display rather implausible time profiles. In some cases, attainment shares fall over time and in others they rise very sharply, displaying increases of over 10 or even 15 points over a 5-year interval that are virtually impossible.

Figure 8: Evolution of upper secondary attainment levels in selected countries according to B&L (2013), % of the 25+ population
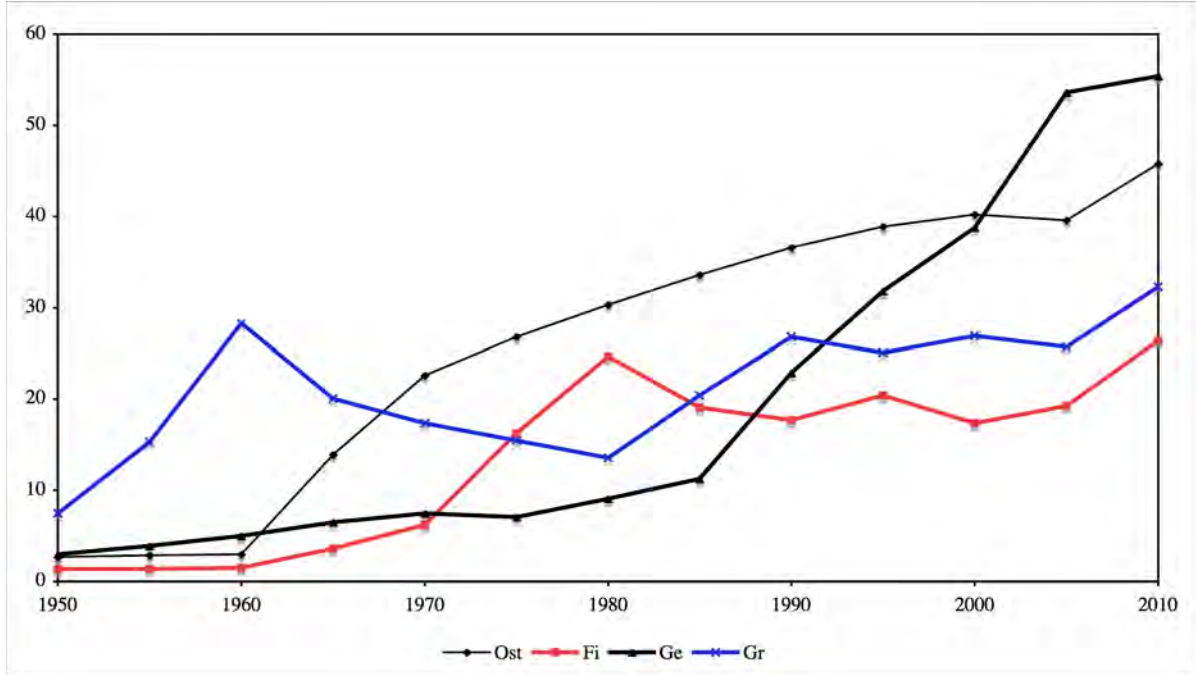
4. Measuring data quality: SUR estimates of reliability ratios

In this section we will construct an indicator of the quality of the different schooling series using D&D's (2006) extension of the procedure suggested by Krueger and Lindhal (K&L, 2001). As K&L note, the information content of a noisy proxy for a variable of interest can be measured by its *reliability ratio,* defined as the ratio of signal to signal plus measurement noise in the data. When several noisy measures of the same magnitude are available, estimates of their respective reliability ratios can be obtained by regressing these variables on each other. Under certain assumptions, the coefficients obtained in this manner can be used to approximate the bias induced by measurement error (which will be a decreasing function of the reliability ratio) and to obtain consistent estimates of the parameters of interest in growth regressions.

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 ε_1 is an iid disturbance with zero mean and uncorrelated with H. The reliability ratio of this series (r_1) is defined as

$$(1) \; r_1 \equiv \frac{var\,H}{var\,P_1} = \frac{var\,H}{var\,H + var\,\varepsilon_1}$$

Assume now that in addition to P_1 we have a second imperfect measure of human capital, $P_2 = H + \varepsilon_2$, where ε_2 is also iid noise. Then, the covariance between P_1 and P_2 can be used to obtain an estimate of the variance of H whenever the measurement error terms ε_1 and ε_2 are uncorrelated. Under this assumption, r_1 can be estimated by

$$(2) \quad \hat{r}_1 = \frac{cov\,(P_1, P_2)}{var\,P_1}$$

which happens 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 $P_2 = c + r_1 P_1$.[11] It must be noted, 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, $\hat{r}_1$ will overestimate the reliability ratio and hence understate the extent of the attenuation bias induced by measurement error.

D&D (2006) build on this approach by exploiting the availability of a number of alternative human capital series to construct a minimum-variance estimator of the reliability ratio. The desired estimator of the reliability ratio of data set k, known as the *SUR reliability ratio*, can be obtained by estimating as a restricted SUR with a common slope a set of equations in which series k is used to try to explain other series, j, i.e. a system of the form

$$(3)\,P_j = c_{jk} + r_k\,P_k + u_{jk} \quad \text{for } j = 1..., K \text{ and } j \neq k$$

The reliability ratio of Barro and Lee's (2013) data set, for instance, can be estimated by using this series of average years of schooling as the explanatory variable in a set of regressions where the dependent variables are the average years of schooling according to other sources.

The exercise we have just described is repeated for several transformations of average years of schooling. In particular, we estimate reliability ratios for years of schooling measured in levels (H_{it}) and in logs (h_{it}), in first differences (ΔH_{it}) and in annual growth rates (Δh_{it}), and for some of these variables measured in deviations from their respective country means $(H_{it} - H_i\,,\,h_{it} - h_i$ and $\Delta h_{it} - \Delta h_i)$. Notice that the last three expressions in this list correspond to the "within" transformations often used to remove fixed effects. We also estimate all the reliability ratios twice, once with the raw data and a second time after removing period means from the different schooling series.

The results are shown in the two panels of Table 6.[12] The last row of each table shows the average value of the reliability ratio for each type of data transformation (taken across data sets), and the last column displays the average reliability ratio of each data set (taken across data transformations). It should be noted that, while reliability ratios must lie between zero and one, some of the estimates reported in Table 6 fall outside these bounds, suggesting that a positive correlation in error terms across data sets may be inflating our estimates of reliability ratios, especially when the data are used in levels or logs.

---

[11] 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 a bias. Hence, it is almost as if we were regressing the true variable on P_1.

[12] In Appendix 2 we undertake the same exercise for non-OECD countries. We find that estimated reliability ratios are somewhat higher in the non-OECD than in the OECD sample. This may be partly the spurious result of a higher correlation of errors across data sets but may also have something to do with the greater variation of schooling in this sample.

Table 6: SUR estimates of reliability ratios, OECD 21 sample

a. Raw data

	Hit	hit	ΔHit	Δhit	Hit-Hi	hit-hi	Δhit-Δhi	average
B&L 12	0.685	0.621	0.006	0.107	0.706	0.620	0.020	*0.395*
	[0.038]	[0.034]	[0.026]	[0.032]	[0.028]	[0.025]	[0.019]	
C&S	0.897	0.877	0.362	0.694	0.969	0.993	0.362	*0.736*
	[0.027]	[0.025]	[0.086]	[0.061]	[0.023]	[0.021]	[0.034]	
L&S	0.608	0.658	0.323	0.482	0.877	0.803	0.341	*0.585*
	[0.022]	[0.018]	[0.033]	[0.024]	[0.022]	[0.020]	[0.045]	
D&D 12	**1.006**	**1.073**	**0.446**	**0.863**	**0.992**	**1.005**	**0.661**	***0.864***
	[0.023]	[0.020]	[0.054]	[0.048]	[0.017]	[0.015]	[0.108]	
average	*0.799*	*0.807*	*0.284*	*0.536*	*0.886*	*0.855*	*0.346*	*0.645*
Obs.	169	169	148	148	169	169	148	

b. Data in deviations from period means

	Hit	hit	ΔHit	Δhit	Hit-Hi	hit-hi	Δhit-Δhi	average
B&L 12	0.638	0.584	0.014	0.090	0.066	0.166	0.008	*0.224*
	[0.043]	[0.038]	[0.024]	[0.031]	[0.026]	[0.029]	[0.015]	
C&S	0.873	0.844	**0.582**	**0.857**	**0.773**	**1.097**	0.263	***0.756***
	[0.031]	[0.027]	[0.094]	[0.063]	[0.107]	[0.054]	[0.048]	
L&S	0.573	0.631	0.285	0.461	0.255	0.456	0.187	*0.407*
	[0.021]	[0.018]	[0.034]	[0.025]	[0.025]	[0.019]	[0.058]	
D&D 12	**0.980**	**1.068**	0.481	0.791	0.481	0.766	**0.302**	*0.696*
	[0.024]	[0.021]	[0.043]	[0.038]	[0.032]	[0.024]	[0.074]	
average	*0.766*	*0.782*	*0.341*	*0.550*	*0.394*	*0.622*	*0.190*	*0.517*
Obs.	169	169	148	148	169	169	148	

Notes:
- Standard errors in brackets below each estimate.
- Data are reported at 5-year intervals except by Cohen and Soto who do it at 10-year intervals. We use linear interpolation (with the data in levels) to complete these series prior to all calculations.
- Panel *a* corresponds to the variables as originally measured. The estimates shown in panel *b* are obtained after removing the corresponding period means. This is done by introducing period dummies in equation (4).
- All equations are estimated using data for 1970-2010, which is the period over which the four series overlap.

Our mean estimate of the reliability ratio in the OECD sample is 0.645 for the raw data and 0.517 after removing period fixed effects. Since these figures are significantly higher than those obtained in our (2006) paper using an earlier generation of schooling data sets (0.386 and 0.335), one encouraging conclusion is that recent studies seem to have succeeded in improving the quality of the data. Even so, a considerable amount of measurement error seems to remain in the data. As is well known, this can generate a substantial downward bias in estimates of the coefficient of schooling in growth equations and production functions, particularly when the data are used in differences or in growth rates. The problem is particularly acute in the case of the B&L data set, which has by far the lowest average reliability ratio both with the raw data and after removing period means, followed by the L&S+ series. Cohen and Soto's and our own series appear to have the highest information content but even in this case the likely bias can be very large in some specifications.

6. Conclusion

In a series of highly influential papers, Barro and Lee have constructed estimates of educational attainment in a broad sample of countries starting from Unesco compilations of census results and using an increasingly sophisticated perpetual inventory procedure to fill in gaps in these data. In a paper written a few years ago (D&D 2006), we pointed out that the versions of B&L's series that were available at the time tended to be rather volatile, presumably as a result of changes in classification criteria, and that this translated into relatively low reliability ratios that alerted of a potentially serious bias toward zero in the estimation of the coefficient of human capital in production functions and growth regressions, particularly when the data were used in differences. In the same paper we constructed an alternative schooling series for an OECD sample that tried to increase the signal-to-noise ratio in the data by introducing previously unused sources to reconstruct plausible time profiles for attainment in each country. A roughly contemporaneous and similarly motivated study by Cohen and Soto (2007) led to similar conclusions and produced a third attainment series that was generally closer to our own than to Barro and Lee's figures.

This paper revisits the issue after a new round of studies that update and improve the available attainment series have been completed. We review the methodology and compare the results of four recent studies (including updates of those cited above and a fourth one due to Lutz, Samir et al) that produce attainment series for different samples of countries during the period between 1960/70 and 2010. We also estimate reliability ratios for each data set using several data transformations that correspond to standard estimation techniques. On the positive side, estimated reliability ratios for the more recent data sets are higher than those for earlier series, suggesting that successive data revisions have succeeded in increasing signal to noise ratios. On the other hand, the results also suggest that the potential attenuation bias continues to be rather high, particularly in differenced specifications. Somewhat surprisingly, even the latest careful revision of B&L's data set has not removed some of its more implausible features. Our estimates of reliability ratios also suggest that this source has the lowest signal-to-noise ratio among the four data sets we compare. We believe these problems have their origin in Barro and Lee's reliance on data from Unesco compilations that are likely to contain a considerable amount of noise. If we are right, the problem cannot be corrected by any improvements in the procedure used to fill-in gaps in the Unesco data, which seems to have been the main focus of B&L's recent work on the issue.

Appendix 1: Details of the construction of some schooling series

1. The basic fill-in procedure

To discuss the details of the fill-in procedure used by Barro and Lee and Cohen and Soto, we need to introduce some notation. For concreteness, for purposes of this section, we will define the adult population as that aged 15 and over and denote it by $L15_t^+$. Following Barro and Lee, we will denote by H_{jt} the number of people aged 15 and over for whom j is the highest level of schooling attained (but not necessarily completed) with $j = 0$ for no schooling, $j = 1$ for primary schooling, $j = 2$ for secondary schooling and $j = 3$ for tertiary or higher education. Dividing H_{jt} by $L15_t^+$ we obtain the fraction of the adult population that has attained level j of schooling, which will be denoted by h_{jt}. Finally, we will use *PRI, SEC* and *HIGH* to denote the primary, secondary and higher enrollment rates.

In the version of the fill-in procedure used in Barro and Lee's earlier papers,[13] the unknown value of H_{jt} is estimated using the known value of the same variable at an earlier (or later) date and data on the size and enrollment rates of the last cohort to enter the adult population. Working in the forward direction, the number of people aged 15 and over with no schooling at time t is approximated by

(1) $\hat{H}_{0t} = H_{0t-5}(1-\hat{\delta}_t) + L_t^{15-19}(1 - PRI_{t-5})$

where L_t^{15-19} is the population aged 15 to 19 at time t and $1-\hat{\delta}$ an estimate of the "survival rate" of the population 15+ between t-5 and t, taking into account migration as well as mortality. Hence, the population 15+ with no schooling at time t will be the sum of two groups: i) those uneducated persons aged 15 or over in t-5 that have survived until t (whose educational level is assumed to remain unchanged) and ii) those persons aged 10 to 14 at t-5 (and hence 15 to 19 at time t) who were not enrolled in primary school five years ago. By the same logic, primary attainment is estimated by

(2) $\hat{H}_{1t} = H_{1t-5}(1-\hat{\delta}_t) + L_t^{15-19}(PRI_{t-5} - SEC_t)$

where the second term captures those persons aged 10 to 14 at t-5 who were enrolled in primary school in t-5 but are not currently enrolled in secondary school (because that would put them in H_{2t} rather than H_{1t}).[14] Similarly, secondary and higher attainment are estimated as

(3) $\hat{H}_{2t} = H_{2t-5}(1-\hat{\delta}_t) + L_t^{15-19} * SEC_{t-5} - L_t^{20-24} * HIGH_t$

(4) $\hat{H}_{3t} = H_{3t-5}(1-\hat{\delta}_t) + L_t^{20-24} * HIGH_t$

---

[13] In Barro and Lee (1993) attainment is only estimated for the population aged 25 and over, so a slightly different version of these equations is used. The basic equations for the 25+ age group remain unchanged in Barro and Lee (1996), although the enrollment rates used in the calculations differ as noted in the text.

[14] Having to work with 5-year population segments (which is often the finest disaggregation available for census data) makes it impossible to replicate exactly the timing of the different cycles. For instance, primary education typically starts before the age of 10 which is assumed here and ends before the age of 14.

where L_t^{20-24} is the population aged 20 to 24 at time t, which may be currently enrolled in higher education.

It is important to note that a single value of $\hat{\delta}_t$ is used in all the above equations, implying a common mortality rate for all cohorts of the adult population and all educational levels. Barro and Lee approximate the "mortality" rate for the overall population aged 15 or over at t-5 by

$$(5) \ \hat{\delta}_t = \frac{L_t^{15-19} + L15_{t-5}^+ - L15_t^+}{L15_{t-5}^+} = \frac{L15_{t-5}^+ - (L15_t^+ - L_t^{15-19})}{L15_{t-5}^+} = \frac{L15_{t-5}^+ - L20_t^+}{L15_{t-5}^+}$$

i.e. by the fraction of the population aged 15 or over in t-5 which does not make it into the population aged 20 or over in t, as a result of either death or migration. Notice that

$$(6) \ 1 - \hat{\delta}_t = 1 - \frac{L_t^{15-19} + L15_{t-5}^+ - L15_t^+}{L15_{t-5}^+} = \frac{L15_{t-5}^+ - L_t^{15-19} - L15_{t-5}^+ + L15_t^+}{L15_{t-5}^+} = \frac{L15_t^+ - L_t^{15-19}}{L15_{t-5}^+}$$

Substituting (6) into (1)

$$\hat{H}_{0t} = H_{0t-5} \left(\frac{L15_t^+ - L_t^{15-19}}{L15_{t-5}^+} \right) + L_t^{15-19}(1 - PRI_{t-5})$$

and dividing through by $L15_t^+$ we obtain the following expression:

$$(7) \quad \begin{aligned} \hat{h}_{0t} &= \frac{\hat{H}_{0t}}{L15_t^+} = \frac{H_{0t-5}}{L15_t^+} \left(\frac{L15_t^+ - L_t^{15-19}}{L15_{t-5}^+} \right) + \frac{L_t^{15-19}}{L15_t^+}(1 - PRI_{t-5}) \\ &= \frac{H_{0t-5}}{L15_{t-5}^+} \frac{L15_{t-5}^+}{L15_t^+} \left(\frac{L15_t^+ - L_t^{15-19}}{L15_{t-5}^+} \right) + \frac{L_t^{15-19}}{L15_t^+}(1 - PRI_{t-5}) \\ &= h_{0t-5} \left(\frac{L15_t^+ - L_t^{15-19}}{L15_t^+} \right) + \frac{L_t^{15-19}}{L15_t^+}(1 - PRI_{t-5}) \\ &= h_{0t-5} \left(1 - \frac{L_t^{15-19}}{L15_t^+} \right) + \frac{L_t^{15-19}}{L15_t^+}(1 - PRI_{t-5}) \end{aligned}$$

Hence, h_{0t} can be estimated as a weighted average of h_{0t-5} and the non-educated fraction of the latest 5-year cohort to enter the adult population, with weights that reflect the shares in the adult population at time t of its two components: those who were already adults at t-5 and new entrants into this group. Similar expressions can be derived for h_{1t}, h_{2t} and h_{3t} by the same procedure.

Barro and Lee (1993) estimate *PRI, SEC* and *HIGH* using data on gross enrollment rates. This variable is defined as the ratio between the total number of students of all ages enrolled in a given educational cycle and the population who, according to national custom or law, "should" be enrolled in that cycle. Barro and Lee (1996) note that the use of gross enrollment ratios will overstate achievement if a significant number of students repeat grades or go in and out of school. To avoid this problem, they use data on net enrollment rates, constructed in the same way as gross enrollments but including in the numerator only those students who belong to the age group that should theoretically be enrolled in the cycle of interest. This practice, however, tends to understate achievement if a significant number

of students are early or late entrants in a given cycle, as is common in many developing countries. For this reason, Barro and Lee (2000) use gross enrollment rates adjusted by repeaters, i.e. try to take into account students of all ages while excluding repeaters.[15]

As we have written them, equations (1)-(4) can be used to make *forward projections*, i.e. to estimate attainment in a given year on the basis of data for earlier years. These equations can also be solved for H_{jt-5} as a function of H_{jt} to write them in a way that can be used to construct *backward projections* for the years before a known census. When a missing observation is surrounded by census data on both sides, attainment can be estimated by either forward or backward projections or by interpolation. As suggested by the results of an accuracy test based on a sample of 30 countries for which relatively complete census data are available, Barro and Lee (1993) choose to fill such cells using a weighted average of the forward projection and a linear interpolation between census data with weights of 0.40 and 0.60. While it is not clearly stated in the relevant papers, it seems likely that this procedure has been maintained in the 1996 and 2001 updates of the B&L data set.

2. Refinements of the fill-in procedure

As Barro and Lee already noted in their first paper (B&L, 1993), their perpetual inventory fill-in procedure is only an approximation and may produce inaccurate results if survival probabilities vary systematically with educational attainment, as seems likely. Cohen and Soto (2007) note that Barro and Lee's perpetual inventory procedure also implicitly ignores the fact that mortality rates certainly depend on age. And since age is correlated with attainment in most countries, this assumption generates a bias that can be quite significant. In particular, when younger generations are more educated than older ones, forward estimates using B&L's perpetual inventory procedure will underestimate attainment, while backward estimates will overestimate it. To see why, refer to equations (1)-(4). As we have already noted, these equations apply a single survival rate to all the different attainment categories of the adult population. If higher attainment groups are younger on average than the rest of the population, B&L's methodology will underestimate their survival rate and therefore their size at time *t* and the opposite will be true for lower attainment groups. As a result, forward estimates of average schooling will be biased downward (and backward estimates will be biased upward).

To avoid this problem, C&S (2007) make full use of the available information on educational attainment by age group. Instead of lumping together the entire population 15+, they work with the attainment level of each 5-year age group of the adult population, measured directly in terms of average years of schooling rather than attainment shares. Denoting by ys_t^a the average years of schooling of the *a*-th 5-year age bracket and by l_t^a the observed weight of this group in the population 15+ at time *t*, the average years of schooling of the 15+ population can be estimated by

(8) $\widehat{ys}_t = \sum_a l_t^a \widehat{ys}_t^a$

---

[15] In the case of higher education, they use gross enrollment rates for lack of repetition data.

Forward and backward projections are then constructed by combining equation (8) with the assumption that, once the age of 25 has been reached, the average attainment of each cohort remains constant over time. Forward and backward projections are obtained by assuming respectively that

$$(9) \quad \widehat{ys}_t^a = ys_{t-5}^{a-1} \quad \text{and} \quad (10) \quad \widehat{ys}_t^a = ys_{t+5}^{a+1}$$

that is, that the average attainment at time t of the age group, say, 55-59, can be approximated by that of the age group 50-54 at t-5 or by that of the age group 60-64 at time t+5. Notice that this procedure implicitly incorporates different survival rates for the different age subgroups of the adult population, as it uses their observed shares in the total population at time t to calculate average schooling in equation (8). As in Barro and Lee's early papers, however, it is still implicitly assumed that survival rates are independent of educational attainment (within each age group). As Cohen and Soto note, another assumption implicit in (9) and (10) is that the schooling level of immigrants or emigrants is the same as that of the rest of the population. If a country receives a net migratory inflow and new entrants are less educated than natives, the forward projection procedure will overstate the level of schooling (because the procedure will not take into account that recent new entrants with below average education will drive ys_t^a below ys_{t-5}^{a-1}) while the backward projection will understate it (because, depending on the age of the new entrants, it may assume that they were already in the relevant population five years ago when this is not the case).

Equations (9) and (10) cannot be used to estimate the attainment level of the age groups at both ends of the age distribution because i) a significant fraction of new entrants into the 15+ population have been enrolled in school between t-5 and t, which will have changed their attainment level, and ii) the oldest group in the age breakdown (typically 65+) combines several 5-year age segments that are not observed independently. To estimate the attainment level of these groups, C&S rely on long series of enrollment data, which they adjust for repeaters and drop-outs to estimate net intake ratios, i.e. the ratio of the number of new entrants into the first course of each educational cycle and the population of the theoretical starting age. These ratios are then used to estimate the fraction of the cohort of interest that entered each educational level in its youth, which makes it possible to calculate its average attainment later on.

B&L (2013) essentially adopt Cohen and Soto's refined procedure for the construction of forward and backward projections, although working with attainment shares, h, rather than average years of schooling. Seeking to improve the accuracy of their projections even further, moreover, they allow survival rates to vary not only with age but also with the level of schooling, although only for the population aged 65 and over (which, according to them, is the only segment of the population for which the assumption of uniform survival rates across educational levels does not hold reasonably well in practice). Hence, B&L's counterparts of the equations that describe the assumptions of forward and backward projections are of the same form as C&S's equations (9) and (10)

$$(9') \quad \hat{h}_{jt}^a = h_{jt-5}^{a-1} \quad \text{and} \quad (10') \quad \hat{h}_{jt}^a = h_{jt+5}^{a+1}$$

for the population 25-64 in the case of (9') and 25-59 in the case of (10'). For the oldest and youngest cohorts, things are a bit more complicated. For the younger age groups, attainment at t is estimated

using the attainment of the same age group at *t-5* and data on enrollments. For example, the forward projection of attainment for the 15-19 age group at time *t* is constructed as

$$(11)\ \hat{h}_{jt}^{15-19} = h_{jt-5}^{15-19} + \Delta enroll_{jt}^{15-19}$$

i.e. by assuming that the observed attainment rate of the age group of interest has increased between *t-5* and *t* by the same amount as the corresponding enrollment ratio.

For the oldest cohorts, the forward projection formula becomes

$$(12)\ \hat{h}_{jt}^{a} = h_{jt-5}^{a-1}(1-\hat{\rho}_j)$$

where $1-\hat{\rho}_j$ is an estimate of the relative survival rate[16] over five years of the population 65+ for which *j* is the highest level of schooling attained. In practice, *j* ranges over only two categories: *H* for highly educated people (with secondary attainment or better) and *L* for low education (no schooling or primary education). The relative survival rate varies only across broad groups of countries but not across age subgroups of the 65+ population. It is estimated separately for OECD and non-OECD countries by using available census data to run a series of regressions of the form

$$(13)\ h_{ot}^{a} = (1-\rho_L)h_{ot-5}^{a-1},\quad h_{1t}^{70+} = (1-\rho_L)h_{1t-5}^{65+},\quad h_{2t}^{a} = (1-\rho_H)h_{2t-5}^{a-1}\quad h_{3t}^{a} = (1-\rho_H)h_{3t-5}^{a-1}\quad \text{for } a = \text{70-4 and 75-9}$$

that yield rather similar estimates of relative survival rates for developed and less developed countries.[17]

Proceeding as in their 1993 paper, Barro and Lee (2013) conduct an accuracy check using those countries for which they have reasonably complete census data in order to determine the "optimal" way to fill in empty cells by combining backward and forward estimates when both are available. To determine the optimal weights of these two variables, they regress observed attainment shares for a given year on forward and backward estimates of the same variable (i.e. on lagged and led census observations). In this occasion, however, they do not include the interpolation as a regressor along with the forward and backward estimates and estimate separate weights for advanced and

---

[16] The relative survival rate of attainment level *j* is defined as the ratio between the survival rate of that group and that of the entire population of interest. Since *h* is a proportion rather than the actual number of people, $1-\hat{\rho}_j$ must be a relative survival rate, not an absolute one. That is, if we have

$$H_{jt}^{a} = H_{jt-5}^{a-1}(1-\delta_j^{a-1})$$

and

$$L_t^{a} = L_{t-5}^{a-1}(1-\delta^{a-1})$$

where L_t^a is the total population of the *a*-th age group at time *t* and δ^{a-1} the observed survival rate for this group, then it must be the case that

$$h_{jt}^{a} = \frac{H_{jt}^{a}}{L_t^{a}} = \frac{H_{jt-5}^{a-1}(1-\delta_j^{a-1})}{L_{t-5}^{a-1}(1-\delta^{a-1})} = h_{jt-5}^{a-1}\frac{1-\delta_j^{a-1}}{1-\delta^{a-1}}$$

so the factor multiplying h_{jt-5}^{a-1} is a relative, not absolute survival rate. To make this clear, we use a different letter for relative rates: ρ rather than δ.

[17] In particular, B&L estimate $\rho_L = 0.966$ and $\rho_H = 1.065$ for OECD countries and $\rho_L = 0.969$ and $\rho_H = 1.068$ for non-OECD ones.

developing countries. The estimated weight of the forward extrapolation is 0.461 for OECD countries and 0.549 for non-OECD countries.

Lutz et al (2007) allow (sex-specific) survival probabilities to depend on educational attainment also for younger age groups. Their backward projections are constructed using an equation of the form

$$(14)\ \hat{H}_{jt-5}^{a-1,s} = \frac{H_{jt}^{a,s}}{1 - \hat{\delta}_{jt}^{a,s}}$$

where the superscript s denotes sex, which in principle imposes fewer restrictions on survival rates than B&L's or C&S's assumptions. In practice, however, limited data availability forces L&S+ to base their estimates of survival rates on fairly restrictive assumptions. On the basis of studies conducted with countries for which good data are available, they estimate that life expectancy at age 16 increases by 1 year as we go from no schooling to primary attainment, and by 2 years as we go from primary to secondary attainment or from secondary to higher education. This pattern is superimposed on country specific estimates of life expectancy at 16 and translated into five-year survival rates using the UN's general model life table (which gives expected mortality rates as a function of life expectancy). (See Lutz et al, pp. 211-12). L&S+ also deviate from B&L in that they do not make use of enrollment data. Instead, they essentially extrapolate attainment patterns across age groups in order to deal with the younger and older segments of the adult population (see Lutz et al, pp. 214-8 for details).

It is clear that the refinements introduced in the fill-in procedure by C&S, B&L and L&S+ do not solve the problems raised by migration. As noted, this will be a significant problem only if migration flows are large and migrants' attainment levels are very different from those of the population. As Lutz et al argue (p. 213), moreover, backward projections will tend to be particularly insensitive to the problem because most migrants are young and therefore drop off the adult population almost immediately as we move back in time.

3. Procedure used for estimating subcategories of attainment

Barro and Lee's fill-in procedure is applied using data on four broad educational categories: no schooling, some primary, some secondary and some higher education. To refine this initial breakdown, these authors make use of the available data on the fraction of the population with each attainment level that has completed the relevant school cycle. That is, let h_j be the fraction of the adult population that has reached level j but not progressed beyond it (with $j = 0$ to 3 for no schooling, primary, secondary and higher education). We can split h_j into two components h_{j1} and h_{j2} capturing respectively the incomplete and complete attainment of level j (as defined by B&L; with the peculiarities noted above in footnote 3) so that $h_j = h_{j1} + h_{j2}$ and define the *completion ratio* for level j as

$$(15)\ c_j = \frac{h_{j2}}{h_j}$$

Completion ratios are available only for some countries and years. To fill in the gaps in such data, Barro and Lee (1993) proceed as follows. (Presumably, this procedure remains unchanged in B&L 1996

and 2001). They first regress the available completion rates (a total of 165 observations) on the lagged (or lead) values of the same variable and on a set of regional dummies. For countries for which completion rates are available for at least one year, the estimated equations are used to fill in missing cells, working either forward or backward. If no observations are available for a country, they use the observed means in the region in which the country is included. In the case of higher education, the data reported in Kaneko (1986) can be used to estimate completion rates for 37 countries around 1980. This ratio is assumed to remain constant over time to extend it to other years. For other countries, regional means are used.

Cohen and Soto (2007) use a similar but simpler procedure. They fill in gaps by assuming that completion rates remain constant over time within each country and assume full completion when no data are available.

In their most recent paper (B&L, 2013), Barro and Lee construct forward or backward extrapolations of completion rates to fill in missing cells in countries where the required disaggregated data are available for at least one year. The forward extrapolation is constructed as follows. For the age groups 25-74, it is assumed that a given cohort's completion rate remains constant over time so that

$$(16) \quad \hat{c}_{jt}^a = c_{jt-5}^{a-1}$$

For younger cohorts, the assumption is that the cross-cohort rate of improvement observed at a given point in time also holds over time. For the 15-19 cohort, for instance, it is assumed that

$$(17) \quad \hat{c}_{jt}^{15-19} = c_{jt-5}^{15-19} * \frac{c_{jt-5}^{20-24}}{c_{jt-5}^{25-29}}$$

ie. that the completion rate for 15-19 age group has improved between *t-5* and *t* in the same proportion as the completion rate observed in *t-5* for 20-24 year olds exceed that for the 25-29 cohort. Finally, for the older cohorts they use a weighted average of the completion rates of the same age group five years ago and of the most recent entering cohort. Thus, the completion rate for the 75-79 cohort is estimated as

$$(18) \quad \hat{c}_{jt}^{75-79} = sh_{t-5}^{70-74} c_{jt-5}^{70-74} + sh_{t-5}^{75-79} c_{jt-5}^{75-79}$$

where sh_{t-5}^{70-74} is the share of the 70-74 group in the total population 70+. As above, when both the forward and the backward extrapolations are available, they are averaged to fill in missing cells. The weights are estimated as above for the case of attainment shares, using regressions of observed completion rates on forward and backward projections. For countries where completion ratios are available only for the population as a whole, cohort-specific rates are estimated by using the typical age profile of completion rates in the region to which the country belongs. When there are no completion data for a given country, they use averages for OECD and non-OECD countries.

Appendix 2: Barro and Lee vs. Cohen and Soto and Lutz, Samir et al outside the OECD21

Unlike D&D, C&S, L&S+ and B&L provide data for a large number of non-OECD countries. Sixty-three countries outside the OECD21 sample used in the text are covered by all three sources. Using this common non-OECD sample, we have fitted distributions to the data in growth rates and estimated SUR reliability ratios for these three data sets. The results are largely consistent with those obtained with the OECD21 sample: the B&L series display the highest volatility, as evidenced by the thicker tails of its estimated distribution in Figure A2.1, and tend to have lower reliability ratios than the other two sources, particularly when we work with the data in differences or growth rates (see Table A2.1).

Figure A2.1: Fitted distribution of the growth rate of years of schooling, different data sets common countries outside the OECD21
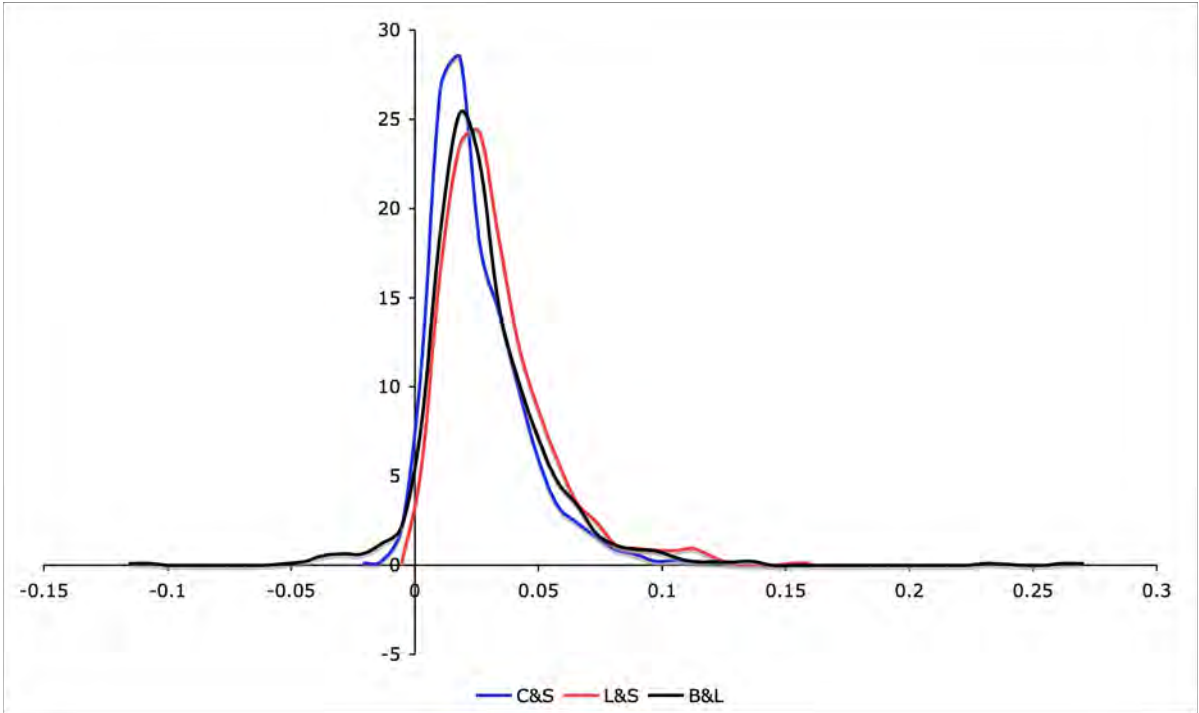
It is worth noting that the estimated reliability ratios are somewhat higher in the non-OECD sample. This is likely to be somewhat misleading, however, because the number of available primary sources that can be drawn upon to construct estimates of educational attainment is probably higher in developed than in underdeveloped countries. As a result, the variation across data sets is likely to be smaller in LDCs, and this will tend to spuriously raise the estimated reliability ratio in a way that will simply reflect a higher correlation of errors across data sets (i.e. an upward bias in the estimated reliability ratio). On the other hand, the result may also have something to do with the fact that the variation of the schooling data is greater in the non-OECD sample. Hence, while we are probably underestimating the amount of noise in this larger sample, it is also likely that the signal will be stronger in it.

Table A2.1: SUR estimates of reliability ratios, non-OECD21 sample

a. Raw data

	Hit	hit	ΔHit	Δhit	Hit-Hi	hit-hi	Δhit-Δhi	average
B&L	1.041	1.015	0.345	0.366	0.984	0.901	0.210	0.695
	[0.011]	[0.015]	[0.025]	[0.027]	[0.011]	[0.016]	[0.022]	
C&S	0.990	0.937	0.519	0.719	1.116	1.045	0.435	0.823
	[0.014]	[0.011]	[0.038]	[0.037]	[0.015]	[0.014]	[0.042]	
L&S	0.798	0.899	0.563	0.578	0.785	0.828	0.429	0.697
	[0.009]	[0.010]	[0.032]	[0.026]	[0.008]	[0.011]	[0.034]	
average	0.943	0.951	0.476	0.554	0.962	0.925	0.358	0.738

b. Data in deviations from period means

	Hit	hit	ΔHit	Δhit	Hit-Hi	hit-hi	Δhit-Δhi	average
B&L	1.041	1.010	0.329	0.316	0.648	0.581	0.112	0.577
	[0.013]	[0.017]	[0.025]	[0.026]	[0.024]	[0.027]	[0.018]	
C&S	0.941	0.894	0.533	0.678	0.657	0.764	0.236	0.672
	[0.015]	[0.011]	[0.039]	[0.035]	[0.032]	[0.025]	[0.036]	
L&S	0.793	0.915	0.544	0.603	0.673	0.722	0.358	0.658
	[0.011]	[0.012]	[0.031]	[0.027]	[0.024]	[0.023]	[0.045]	
average	0.925	0.940	0.469	0.532	0.659	0.689	0.235	0.636

Notes:
- Standard errors in brackets below each estimate.
- Data are reported at 5-year intervals except by Cohen and Soto who do it at 10-year intervals. We use linear interpolation (with the data in levels) to complete these series prior to all calculations.
- Panel *a* corresponds to the variables as originally measured. The estimates shown in panel *b* are obtained after removing the corresponding period means. This is done by introducing period dummies in equation (4).
- All equations are estimated using data for 1970-2010, which is the period over which the four series overlap.

References

Barro, R. and J-W. Lee (1993). "International Comparisons of Educational Attainment." *Journal of Monetary Economics* 32, pp. 363-94.

Barro, R. and J-W. Lee (1994). "Data set for a panel of 138 countries." Mimeo, Harvard University.

Barro, R. and J-W. Lee (1996). "International Measures of Schooling Years and Schooling Quality." *American Economic Review* 86(2), *Papers and Proceedings*, pp. 218-23.

Barro, R. and J-W. Lee (2000). "International data on educational attainment, updates and implications." NBER Working Paper no. 7911. (http://www.nber.org/papers/w7911).

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

Barro, R. and J-W. Lee (2013). "A New Data Set of Educational Attainment in the World, 1950-2010." *Journal of Development Economics* 104, pp. 184-98.

Barro and Lee (2012). "Appendix Notes to A New Data Set of Educational Attainment in the World, 1950-2010." Mimeo. http://www.barrolee.com/

Cohen, D. and M. Soto (2007). "Growth and Human Capital: Good Data, Good Results." *Journal of Economic Growth* 12, pp. 51–76.

de la Fuente, A. and R. Doménech (2002). "Educational attainment in the OECD, 1960-90." CEPR Discussion Paper no. 3390.

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

de la Fuente, A. and R. Doménech (2012). "Educational attainment in the OECD, 1960-2010." Mimeo, Instituto de Análisis Económico, CSIC and BBVA Research. http://ideas.repec.org/p/aub/autbar/911.12.html

Kaneko, M. (1986). "The educational composition of the world's population: a database." Washington DC, the World Bank, Education and Training Department, Report No. EDT 29.

Krueger, A. and M. Lindahl (2001). "Education for growth: why and for whom?" *Journal of Economic Literature* XXXIX, pp. 1101-36.

Kyriacou, G. (1991). "Level and Growth Effects of Human Capital, A Cross-Country Study of the Convergence Hypothesis." Mimeo, NYU.

Lau, L, D. Jamison, and F. Louat (1991). "Education and Productivity in Developing Countries: an Aggregate Production Function Approach." Report no. WPS 612, the World Bank.

Lau, L., S. Bhalla and F. Louat (1991). "Human and physical capital stock in developing countries: construction of data and trends." Draft mimeo, *World Development Report,* World Bank.

Lee, J-W. and R. Barro (2001). "Schooling quality in a cross-section of countries." *Economica* 68, pp. 465-88.

Lutz, W., A. Goujon, K.C. Samir and W. Sanderson (2007). "Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000." *Vienna Yearbook of Population Research 2007*, pp. 193-235. http://www.iiasa.ac.at/Research/POP/edu07/index.html

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, pp. 379-401.

Psacharopoulos, G. and A. Arriagada (1986). "The educational composition of the labour force: an international comparison," *International Labour Review*, 125(5), Sept.-Oct., pp. 561-74.

Samir KC, B. Barakat, A. Goujon, V. Skirbekk , W. Sanderson and W. Lutz (2010). "Projection of populations by level of educational attainment, age, and sex for 120 countries for 2005-2050." *Demographic Research*: Volume 22, Article 15.

UNESCO (various years). *Statistical Yearbook,* Paris.

United Nations (various years). *Demographic Yearbook*, New York.

Appendix 3: Average years of schooling, country profiles

This appendix compares the profiles of estimated years of schooling according to the different sources we review. Figure A1 shows the unweighted average of all the countries in the OECD21 sample we work with. Figures A.2-A.22 compare the different sources for each country in the sample.

**Figure A1: years of schooling in the OECD 21,
unweighted average according to different sources**
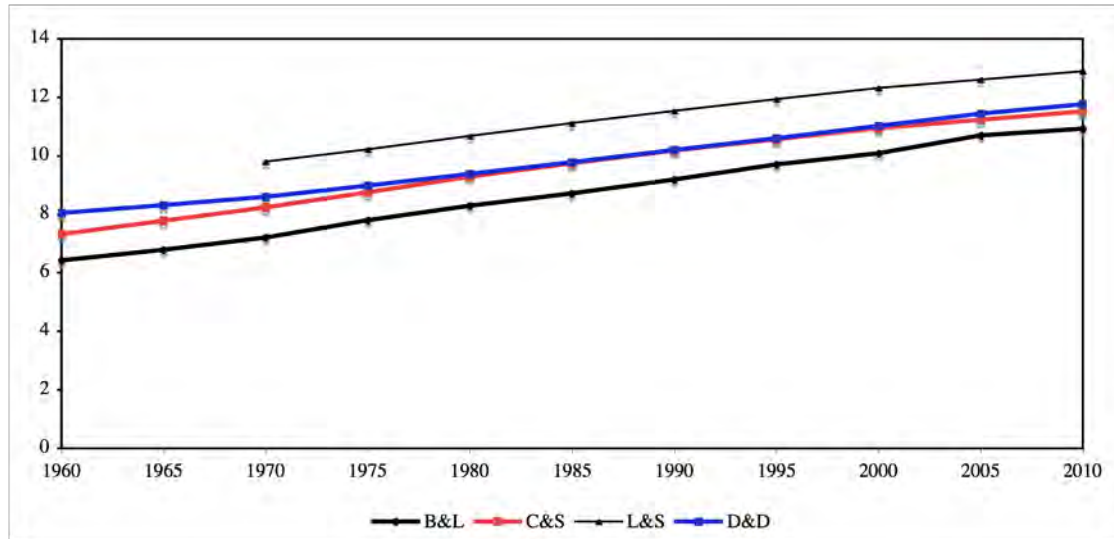
Figure A.2: Australia
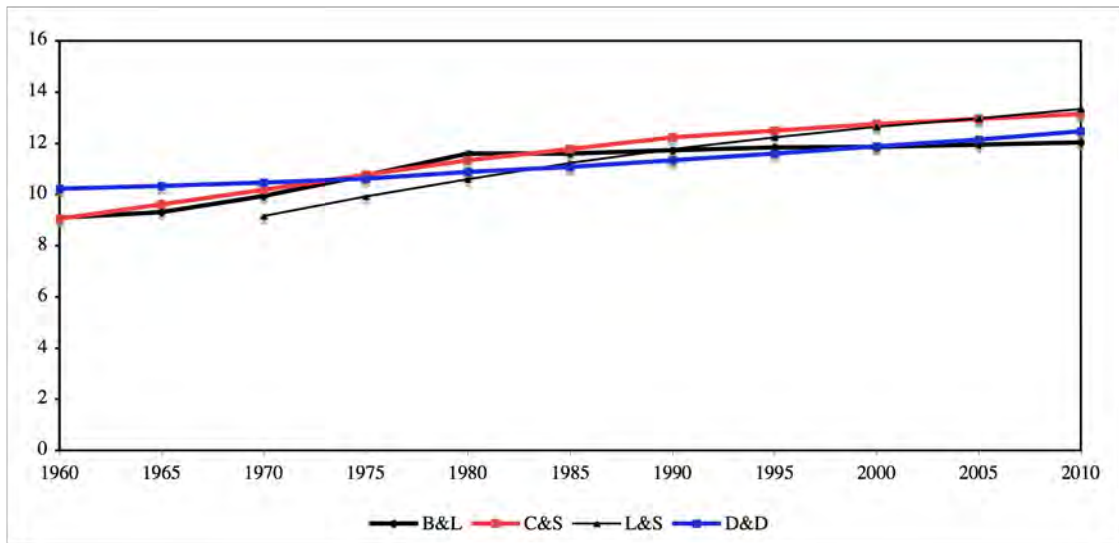
Figure A.3: Austria
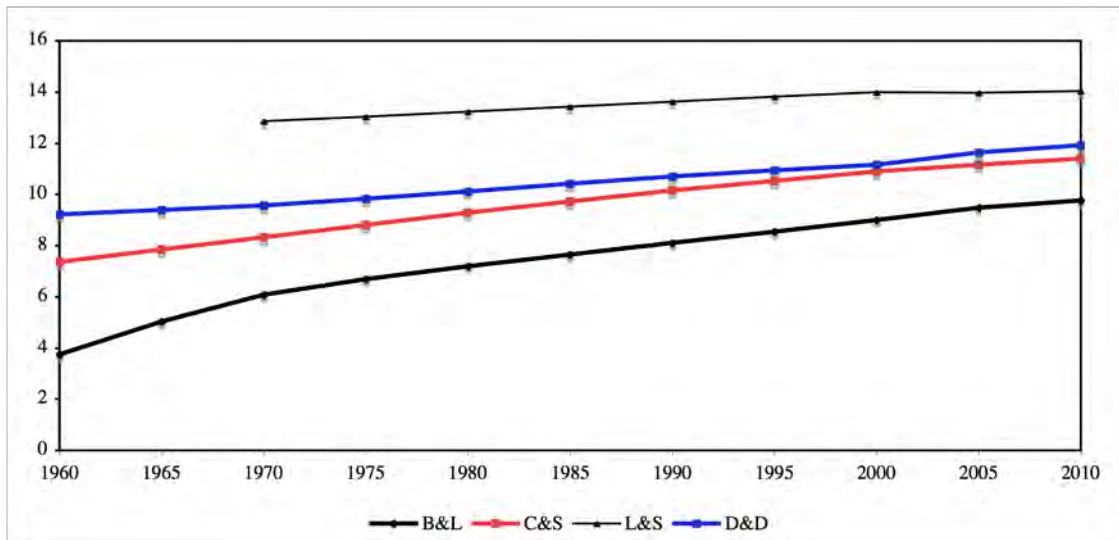
Figure A.4: Belgium
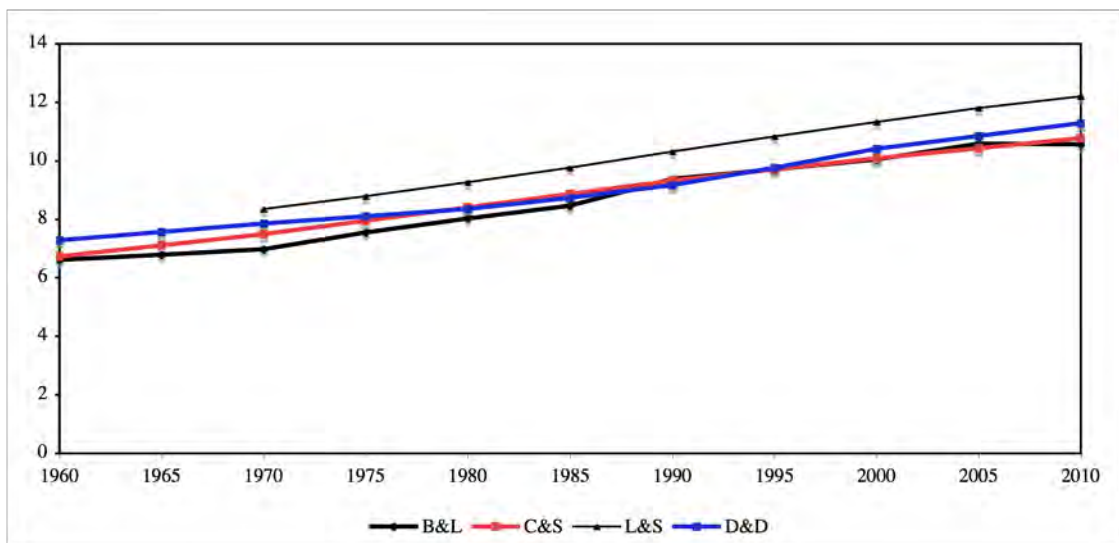
Figure A.5: Canada
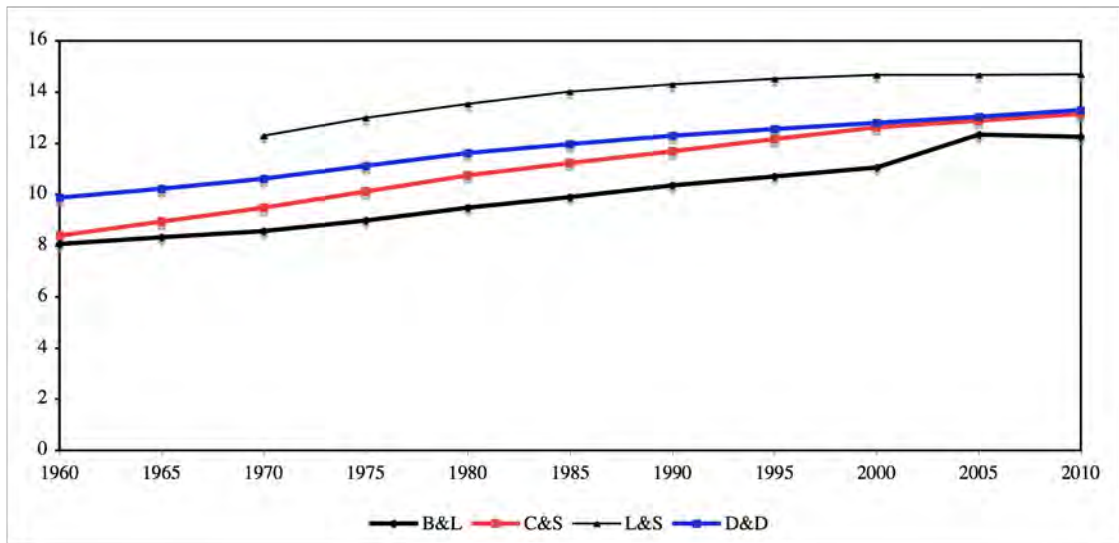
Figure A.6: Denmark
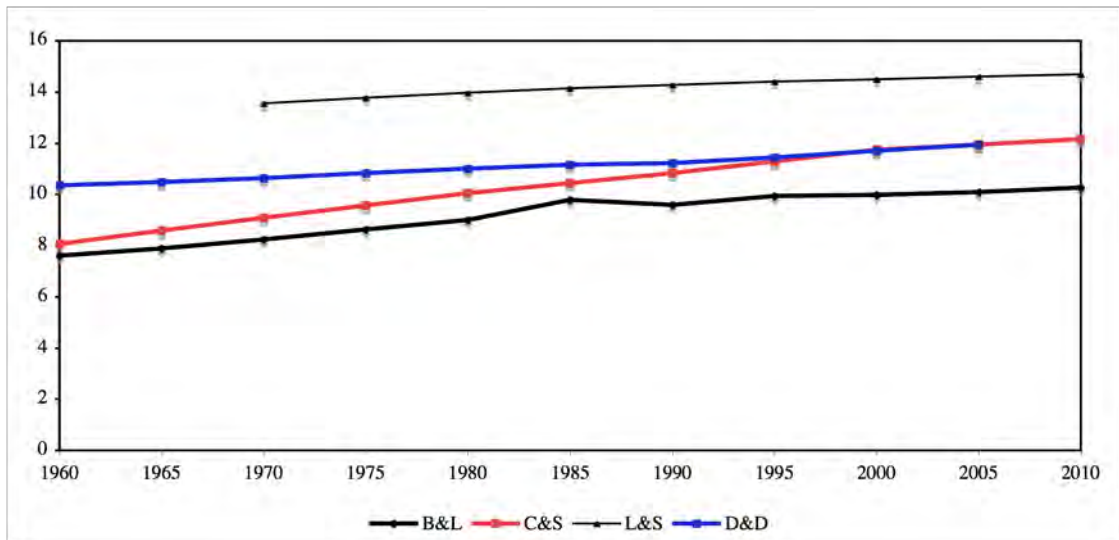
Figure A.7: Finland
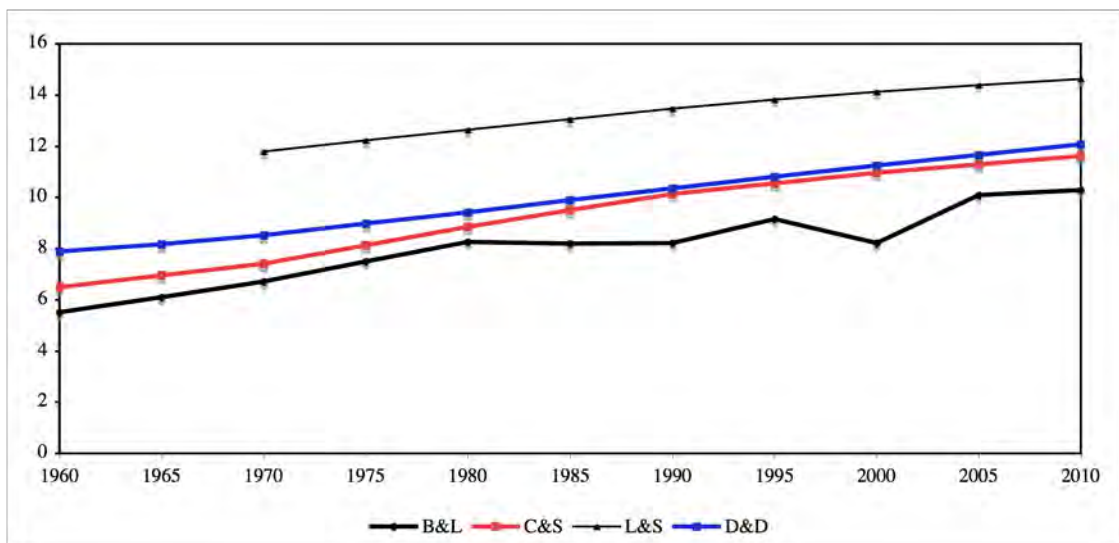
Figure A.8: France
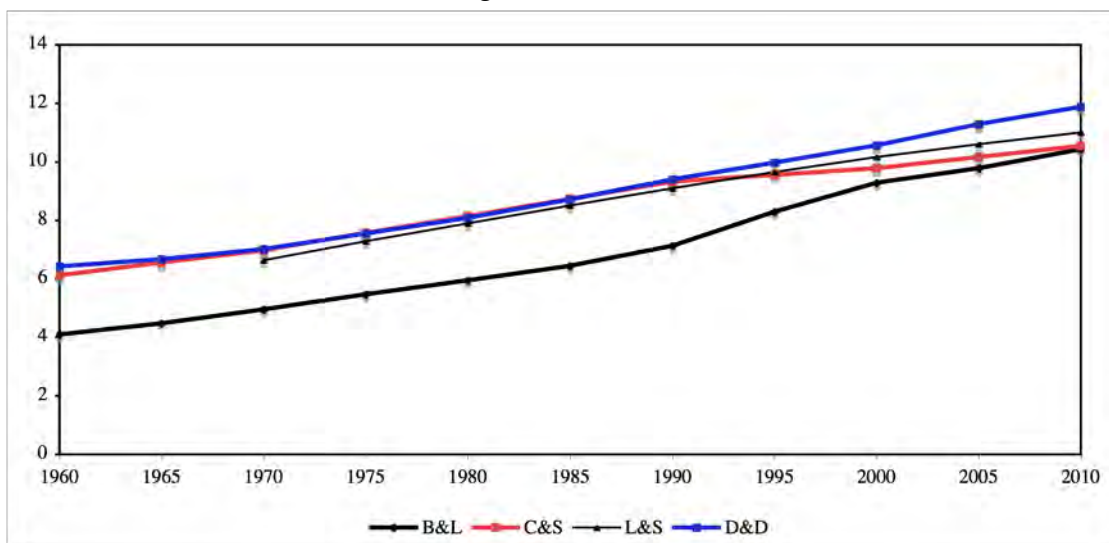
Figure A.9: Germany
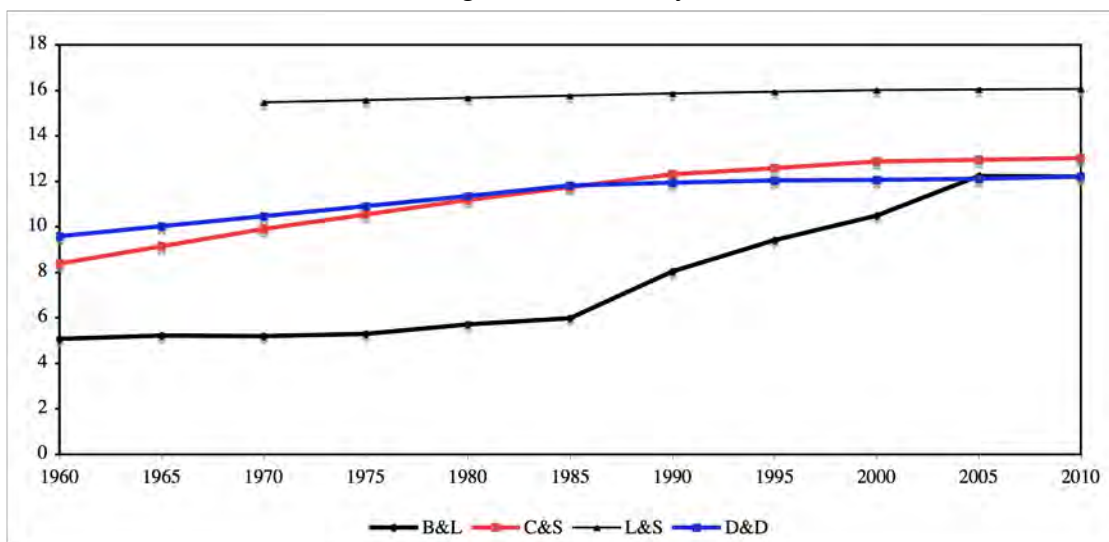
Figure A.10: Greece
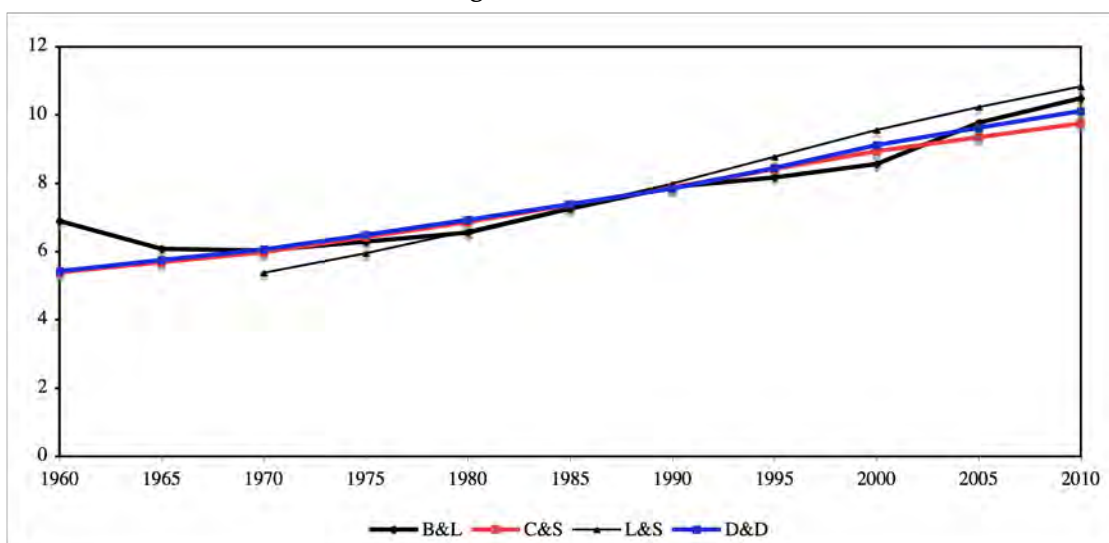
Figure A.11: Ireland
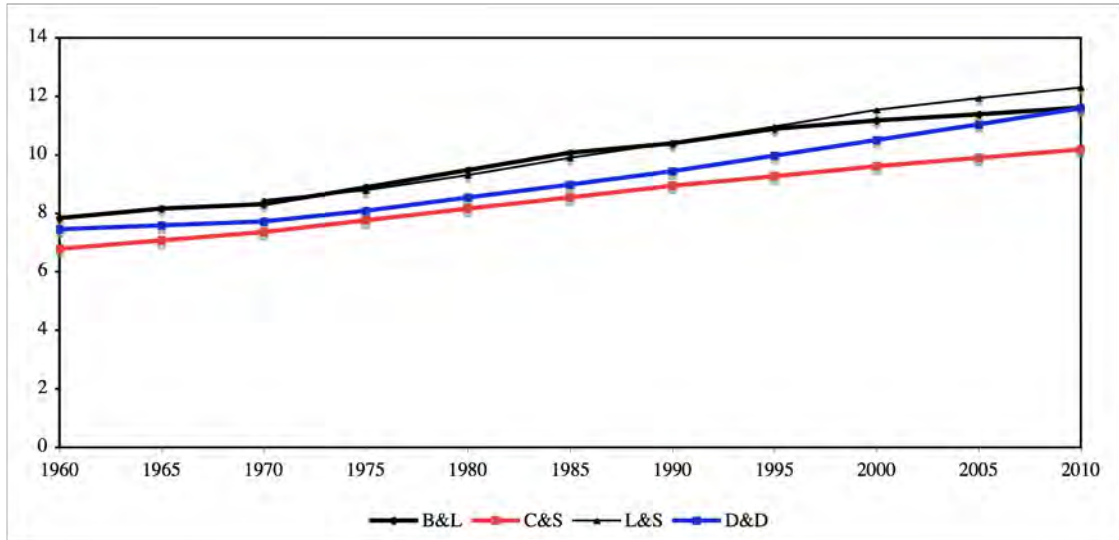
Figure A.12: Italy
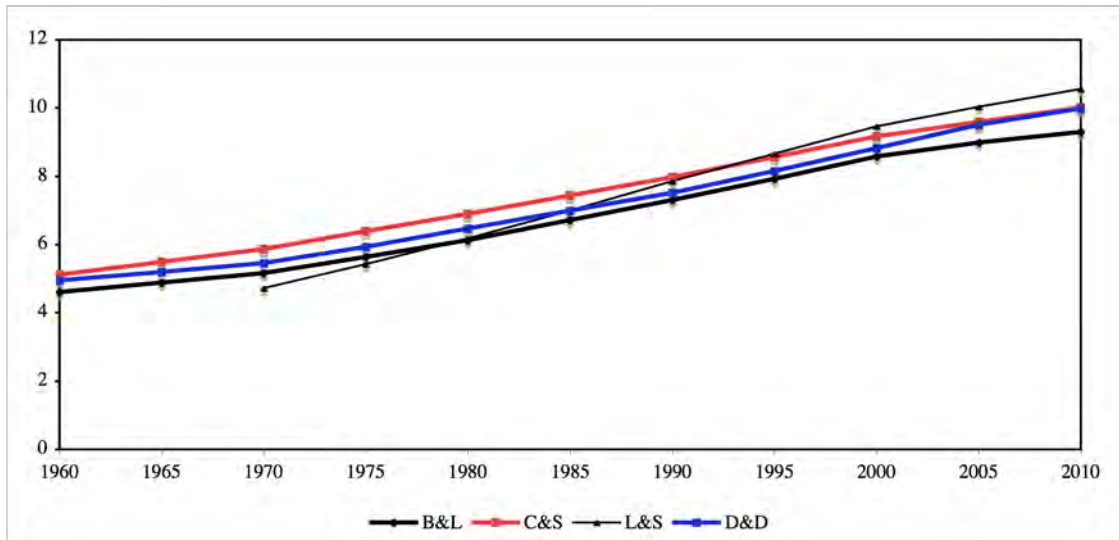
Figure A.13: Japan
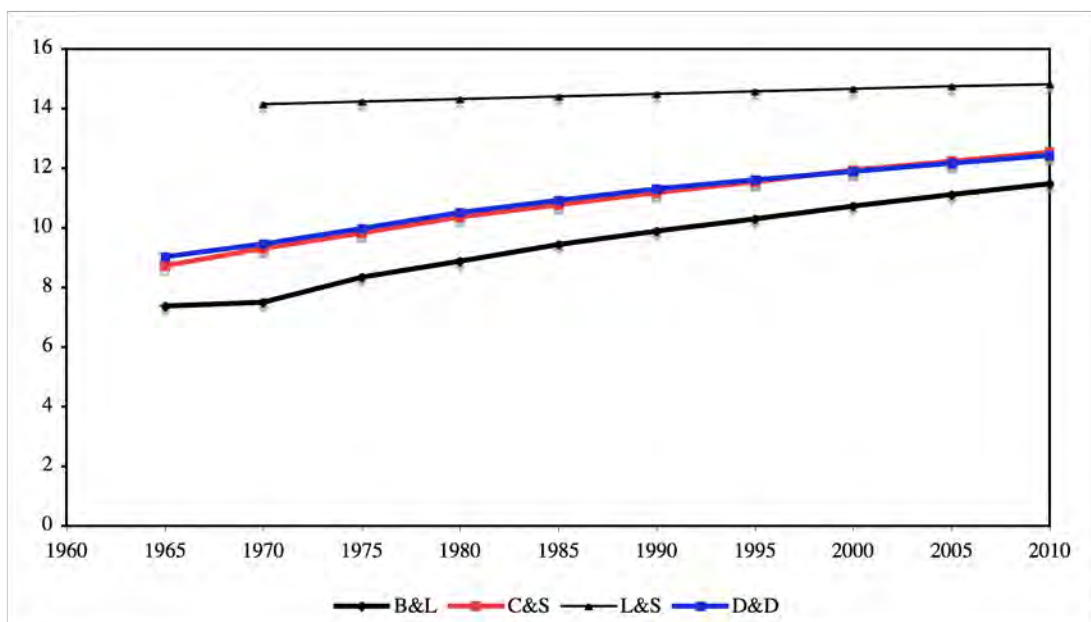
Figure A.14: Netherlands
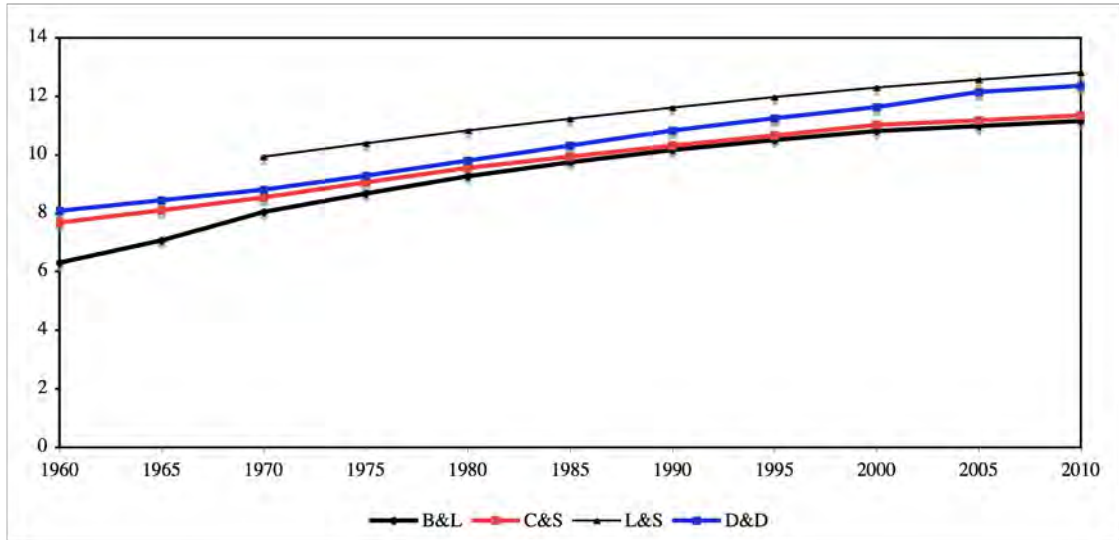
Figure A.15: New Zealand
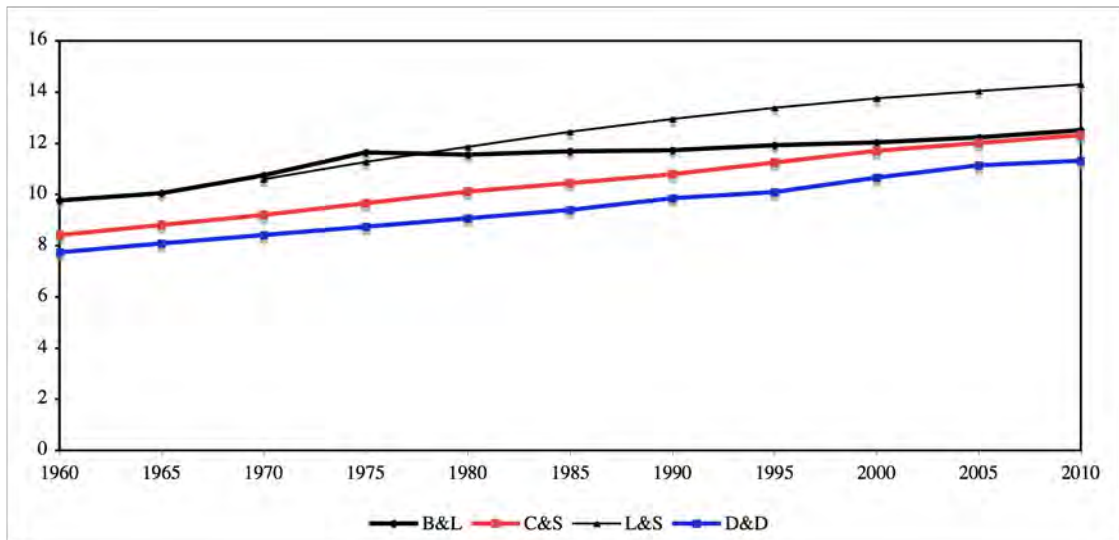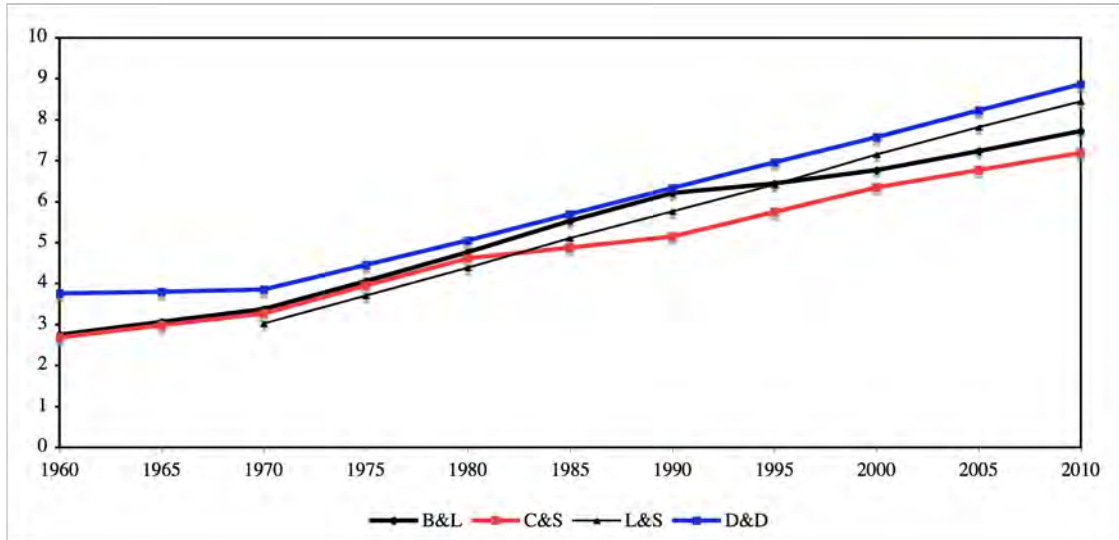

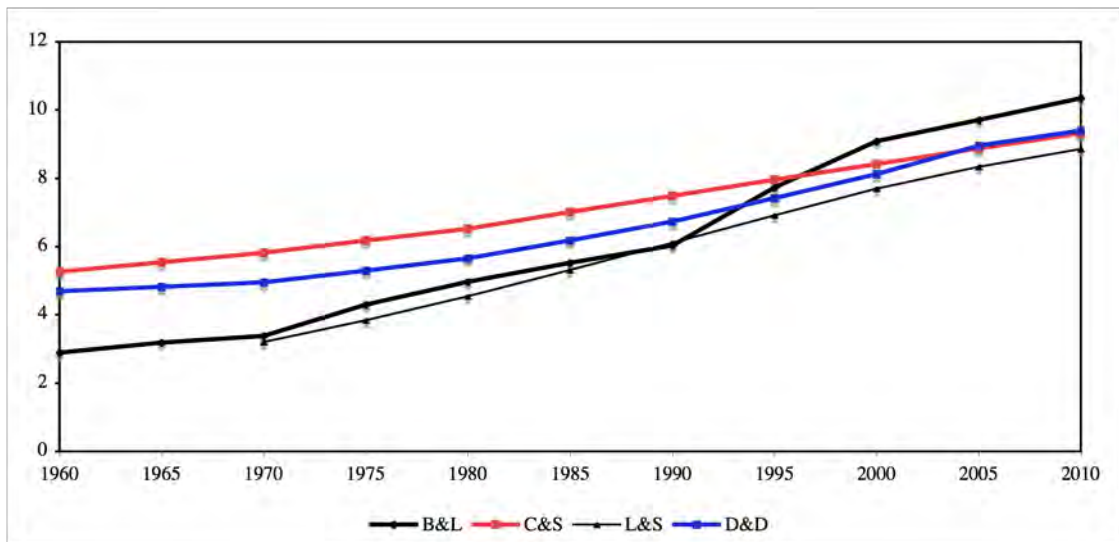## Figure A.16: Norway
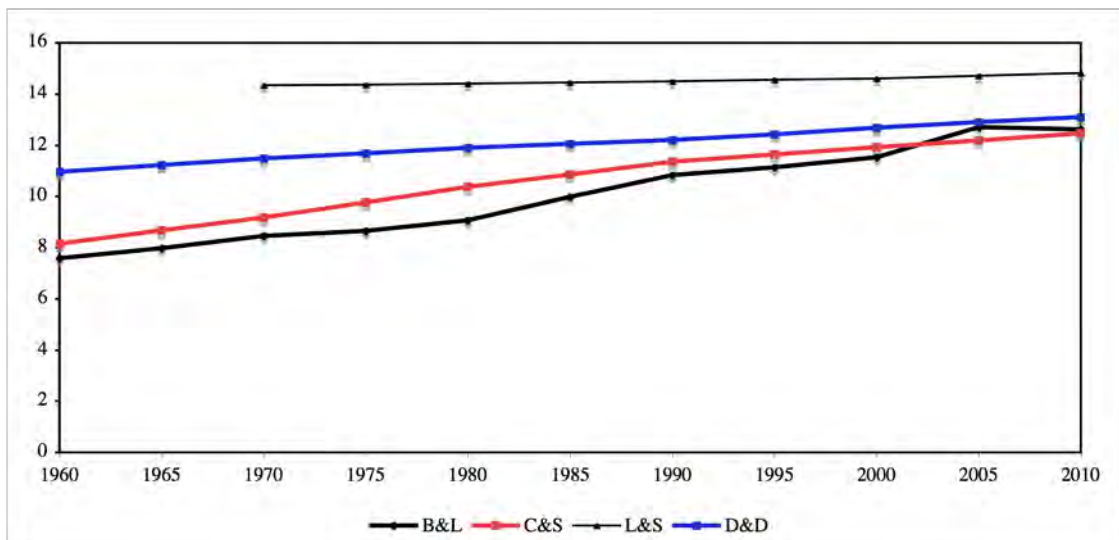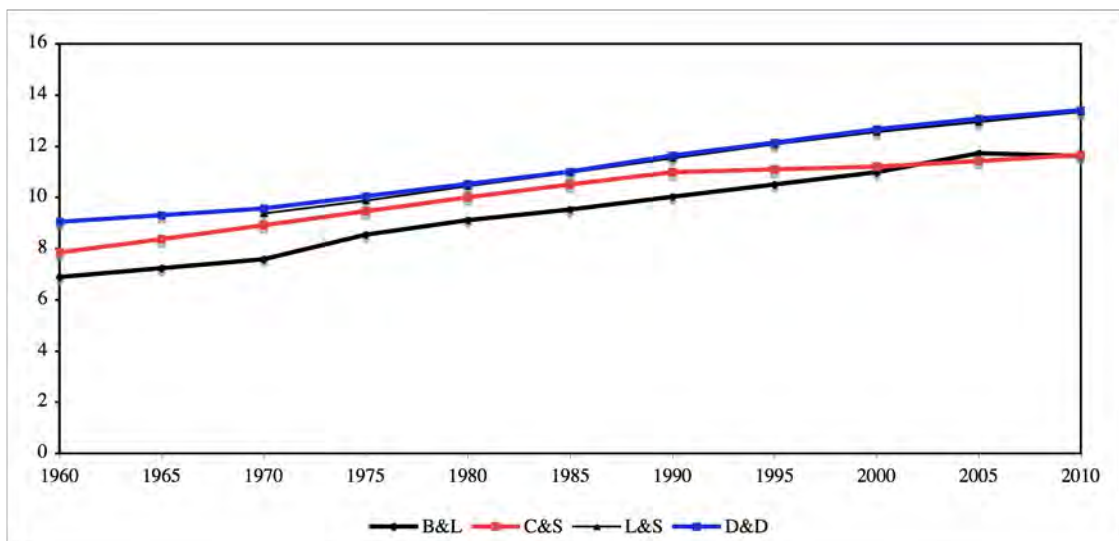
Figure A.17: Portugal

Figure A.18: Spain
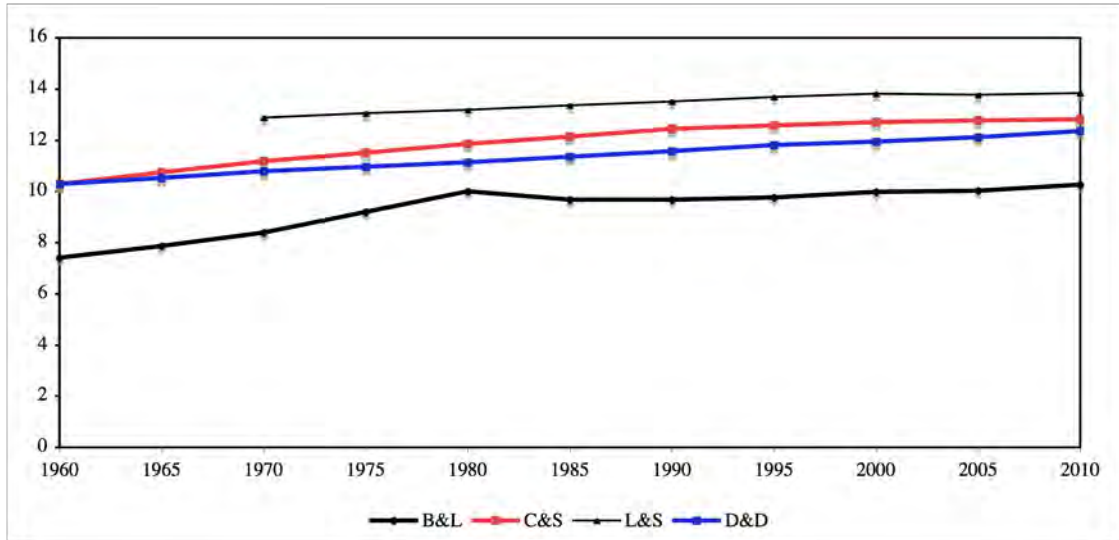

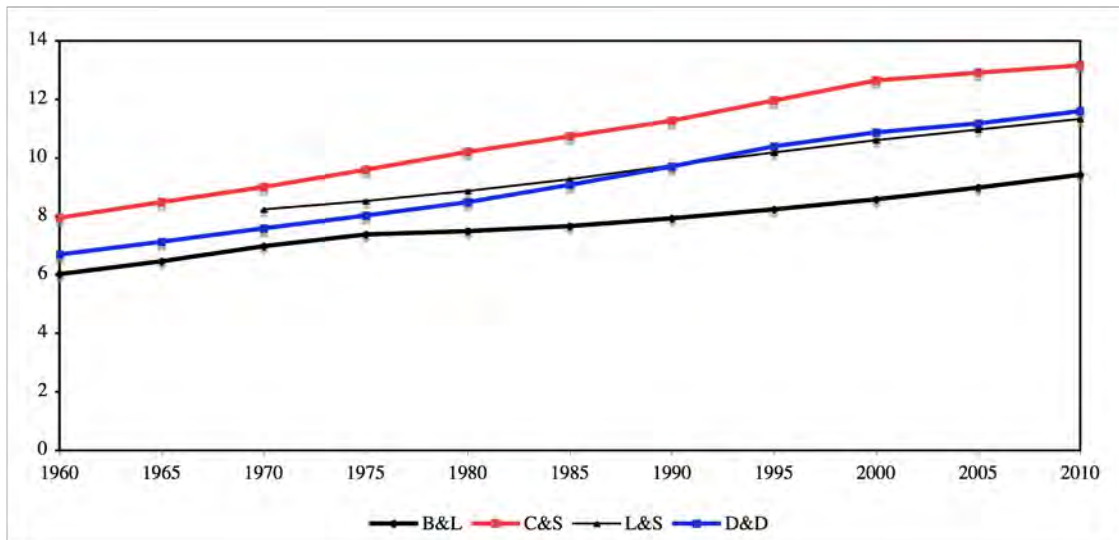**Figure A.19: Sweden**

Figure A.20: Switzerland

Figure A.21: United Kingdom

Figure A.22: United States