UNIVERSITÀ
DEL SALENTO

Quaderni del Dipartimento di Scienze dell'Economia

Education, Earnings and Bias in Ols Estimates: The Role of Birth Order and Family Size

Emanuele GRASSI

Numero E/12
Aprile 2015

Dipartimento di Scienze dell'Economia
Facoltà di Economia
Università del Salento - Lecce
ISSN 2284-0818

Quaderni del Dipartimento di Scienze dell'Economia

I Quaderni intendono favorire la circolazione di studi in versione preliminare, al fine di accelerarne la diffusione e per sollecitare suggerimenti e critiche.

Vengono considerati per la pubblicazione sui Quaderni del Dipartimento solo lavori non ancora pubblicati in versione definitiva su libri o su rivista. Possono proporre lavori per la pubblicazione membri del Dipartimento (anche in collaborazione con studiosi esterni) o studiosi che abbiano presentato il lavoro in oggetto nell'ambito di conferenze o seminari svolti nell'ambito delle attività dipartimentali.

I quaderni si articolano in quattro sezioni: Aziendale, Economica, Giuridica, Matematico Statistica.
I quaderni vengono regolarmente inviati a Istituti, Dipartimenti e Biblioteche a carattere economico e sono inclusi nella serie Research Papers in Economics (REPEC).

L'accesso ai Quaderni è approvato dal Coordinatore scientifico della sezione.

Editore:

Dipartimento di Scienze dell'Economia
Indirizzo mail: direzione.dipeconomia@unisalento.it
Ecotekne - via Monteroni
73100 Lecce

Codice ISSN: 2284-0818

EDUCATION, EARNINGS AND BIAS IN OLS ESTIMATES: THE ROLE OF BIRTH ORDER AND FAMILY SIZE

Emanuele Grassi

Department of Management, Economics, Mathematics and Statistics

University of Salento

Abstract

In this paper I use a sample of 6141 individuals from the 13th wave of the British Household Panel Survey (BHPS) to investigate the relationship between earnings and education using family size and a birth order index as instruments for years of education. Both family size and the birth order index seem to be credible instruments in order to explain the duration of schooling and to interpret the cross section estimates. They allow determining the choice of schooling and then estimating a wage equation to obtain an estimate of the rate of return to education.

I also use information on gender, ethnicity and a set of other variables in order to deal with heterogeneity across individuals, and different family background. I find that the estimated IV coefficients are not dissimilar from the simple OLS estimates, also when a number of controls are introduced in the estimation.

Introduction

Governments and individuals devote part of their (scarce) resources to finance education with the awareness that this can be conceived as a good investment in human capital because of the positive repercussions in terms of self-satisfaction, greater opportunities in the job market, higher earnings and better performances of the economy as a whole (in terms of productivity and growth).

This study is aimed at providing a consistent estimator of the rate of return to education for a cross section of individuals in the presence of the so-called ability bias, measurement error bias and sample selection bias. The theory suggests that in the absence of a good proxy variable able to capture the effect of individual's ability on education, OLS estimates produce bias results. It is therefore necessary to find an exogenous regressor for years of education, and this paper provides reasoning in support of the use of family size and birth order index as instrumental variables.

The paper is organized as follows. In the second section I present a brief overview of the wide literature on this field of research and discuss the economic model and some important implications. Theoretical and methodological progress has also stimulated a stream of empirical work and I devote part of this section to present some of these results. I then proceed to describe the econometric model and the methodology applied focusing on particular issues arising from measurements errors, endogeneity of years of education and heterogeneity across individuals. This section also explains how data has been manipulated and how some variables have been obtained. There follows a section in which I summarize the econometric results and present the conclusions.

Literature review and empirical evidence

In the past fifty years many economists have been involved in a passionate debate aiming to clarify on the one hand the role of human capital accumulation for an economy (in the attempt to explain the determinants of growth rates and productivity paths) and, on the other hand, the relationship between individual's earnings and education. This section explores the theoretical framework in which the second field of research lies.

2.1 Progresses in 50 years of research

The estimation of the return to education is an attempt to partially explain earning differentials among people. The underlying idea is that there is a causal link between education and earnings such that the human capital accumulation increases the probability of getting higher earnings through the build-up of competencies, skills and knowledge.

The natural starting point is the Mincer's (1974) earnings function in which years of education and experience provide an explanation of (log) earnings[1]:

$$\log Y_i = \beta_1 + \beta_2 S_i + \beta_3 X_i + \beta_4 X_i^2 + e_i \qquad (1)$$

In the above equation *Y* is labour income, *S* represents years of education, *X* is experience, *e* is the error term and *i* is an index for individuals. β_2 is commonly referred as the (average) rate of return to education. Equation (1) takes several hypotheses as given. First of all, the functional form is assumed to be linear for schooling and quadratic for experience. While the linearity of years of education is widely assumed, there is much more uncertainty on the experience term. Using three different datasets, Heckman and Polachek (1974) confirmed the Mincer's specification but with some differences in the experience term[2], while Murphy and Welch (1990) suggest a more accurate choice of the functional form (different from the quadratic) and they provide evidence on how a third (or even fourth) order polynomial approximation of the experience term is preferable[3].

Another strong assumption in (1) is that years of education are assumed to be exogenous. In order to show the implications of this hypothesis let assume the following variation of (1):

$$\log Y_i = \beta_1 + \beta_2 S_i + \beta_3 \theta_i + e_i \qquad (1a)$$

In (1a) θ_i represents the so-called ability factor which directly affects earnings. This term can be explained by arguing that more ability is associated with higher productivity and therefore higher earnings. But people with higher ability may perform better also at school

[1] See Card (1999), p.1804.

[2] In particular, the authors find that for different datasets the quadratic form which captures the effect of experience can be replaced by a natural logarithm form.

[3] After having tested several hypotheses, the functional form chosen for this study is linear for schooling and quadratic for the experience term.

and they might choose to take more education[4]. Thus ability affects the choice of S_i as well, and it appears in the following equation:

$$S_i = \gamma_1 + \gamma_2 \theta_i + v_i \tag{2}$$

Let b_2 be the OLS estimator of β_2; the expected value will be[5]:

$$\mathrm{E}(b_2) = \beta_2 + \gamma_2\, cov(\theta S)/var(S) \tag{3}$$

It is then readily shown that a problem of correlation between schooling and earnings arises. As a result, a standard OLS estimate produces an upward biased coefficient of the schooling term, and the attempt to explain higher earnings might be misleading. Although the existence of this problem is not questionable, the magnitude of this bias might be small and/or partially offset by measurement error bias which works in the opposite direction.

This issue has been addressed by many authors using different approaches. Griliches (1977) offers a way to measure the ability factor including IQ[6] test scores in the regressions, and his key finding is that ability can explain a little of the variation in earnings[7]. On the same line of research, but with opposite results, McKinley and Neumark (1995) analyse data from the National Longitudinal Survey of Youth in order to isolate the effect of ability[8]. Based on a sample size of 1592 individuals aged 14-22, they find that if the ability factor is ignored then the OLS estimate is upward biased of about 40%.

Another way to understand the magnitude of the bias in the OLS estimates is to use samples of genetically equivalent twins. Twins based studies have the main advantage that the effect of ability and of family background tend to vanish. Assuming that identical twins[9] cannot differ sensitively in their ability, a within family estimate of the rate of return to education is automatically cleaned from the ability bias[10]. Furthermore, if family background proxies are available from the dataset it is easy to control for between family variability.

Following this approach Miller et al. (1995) carried out a study on a sample of 1170 twin pairs. The authors compare the results of three different estimations. A first one is based on a sub sample of identical twins and is aimed to provide an estimation which is not affected by ability and family background biases. A second estimation is run on a sub sample including not genetically identical twins and should provide an estimation which only suffers the ability bias. In the end, an overall OLS regression provides an estimator affected by both kinds of biases. Their finding is that the upward bias of the rate of return obtained with OLS is in the range of 40-50%. An important remark is that their sample lacks of direct information on labour income but contains data on individual's occupation. Earnings are then approximated with the average labour income of each category. This seems to be a large assumption on the dataset and perhaps the results should be read with care. Card (1999) suggests this interpretation and prefers estimations *à la* Ashenfelter-Rouse (1998) in which a

[4] Abler individuals have an incentive to acquire more education. See, among others, Ashenfelter and Rouse (1998), p.253.

[5] See Griliches (1977) and Card (1999) for further details.

[6] See Hause (1972) for a discussion of the use of IQ test scores as a proxy for ability.

[7] The author finds that the upward bias is only 0.01.

[8] The authors use also test scores of the Armed Service Vocational Aptitude Battery which cover, among the others, mathematics knowledge, arithmetic reasoning and numerical operations.

[9] Identical twins (or monozygotic twins) are those who share the same genetic composition because they come from the same fertilized egg.

[10] See Miller et al., (1995) for a discussion on this topic.

10-15% bias is found. In this study the authors collect data of 700 genetically identical twins and find that a cross sectional estimator of the return to schooling is generally upward biased.

The literature has also investigated in deep the role of family background variables (such as parent's or brothers' education). In particular, some authors have estimated models in which family background variables appear as regressor in the wage equation. In the attempt to explain earnings inequality among Panamanian males, Heckman and Hotz include parents' education in a standard Mincer's earning equation and find that the OLS estimation of the rate of return to education falls of about one third[11]. Lam and Schoeni (1993) show that the omission of family background variables can lead to a bias estimator of the return to schooling. The authors survey a sample of 40000 Brazilian males aged 30-55 in order to isolate the effect of parents, wife and parents in law's education on labour income. Noticing high returns to schooling in developing countries, the authors provide evidence of the upward bias in OLS estimations when one fails to control for family background characteristics. Hence, both studies support the idea that a simple OLS regression tends to overestimate the effect of education on earnings and omitting family background variables leads to higher and misleading estimations.

Differently from other studies, Maluccio (1998) uses parental education as an instrument for years of education (and not as a control in the wage equation). As pointed out in Card (1999), this approach, relatively to the study in question, leads to more precise estimated coefficients.

More recently, some authors have focused their attention on the research of credible instrumental variables related to institutional features of the school system. The analysis of the supply side of schooling systems moves from the idea that if years of education are related to exogenous variables (such as tuition costs, geographic proximity of schools, teacher's salary, expenditures per pupil or compulsory schooling laws[12]), then an Instrumental Variable approach should give a consistent estimator of the rate of return to schooling[13]. Card (2001) criticises this approach arguing that institutional features might affect the distribution of ability across individuals (and so the problem of endogeneity is not solved), but he also proposes a way to try to overcome this issue. In particular, if a second control variable can be found, and under specific hypotheses[14], then the IV estimator is no more affected by endogeneity bias. Studies based on institutional features often reveal higher IV estimates than the corresponding OLS estimates. Card (1995) uses the proximity to college or university in order to obtain an exogenous determinant of years of education. He finds that this variable is important to understand the choice of continuing education, and since in the study there are evident differences related to heterogeneous family backgrounds, the author considers the interaction between proximity and family background variables as a valid instrument for schooling. The result is still a higher IV estimator, but lower in magnitude. On the same line of research, Maluccio (1998) finds strong evidence of the impact of school proximity to completed education.

2.1.1 The role of birth order on education

In the present study several hypotheses suggested by the economics of the family are taken into account. One of the main results of this branch of the literature tells us that the per capita resources that parents assign to their children's education are decreasing in the family size and may not be equally distributed across siblings. Given a certain endowment devoted

[11] In particular, the authors found high levels of significance of the impact of mother's education on male earnings.

[12] See Card, (2001), pp.1127, 1135.

[13] A formal proof of this statement can easily found in Card, (1999) and (2001).

[14] See Card (2001), p.1140 for analytical details.

to children's educational investments, it is very unlikely that after new birth(s) parents increase proportionally the amount of initial resources[15]. Whereas researchers converge on the sign of the family size effect, it is more difficult to establish precisely the sign and the magnitude of the birth order effect.

Siblings with different birth order may be assigned dissimilar resources for their education and may differ for several reasons. Elder siblings may benefit from greater parents' time spent with them but if there is a sensible age difference between siblings, then younger siblings may take advantage of the time spent either with parents or other siblings (Behrman and Taubman, 1986). We notice also that first born children experience both the 'only' child status and the composite family structure status, widening in this way the spectrum of stimuli. Furthermore it is well known that older children are often required to take more responsibility and this can lead to better educational performances (Booth and Kee, 2005). Other hypotheses predicting differences in siblings' assigned resources move from the idea that younger parents have less experience and there might be a sort of learning-by-doing process which may have positive effects on higher birth order siblings[16].

One of the main problems in this field of research is that the effect of family size and birth order must be identified correctly. Since the birth order has a separate effect on education, and we also want to understand the sign of this effect, we need to be sure that the family size effect is actually purged from any birth order effect. The literature on this topic provides some results achieved following different approaches, such as studies with dummy variables (Black et al., 2005) and relative measures of birth order (Ejrnaes and Portner, 2004). But among different approaches, I choose the one adopted in Booth and Kee (2005). The authors build a composite birth order index which is particularly efficient in capturing the birth order effect independently from the family size effect. Their result is robust to a number of specifications[17] and show that for lower birth order the educational attainment is increasing. In other words, the regression output is a negative coefficient for the birth order that can be interpreted as the resultant of all the hypotheses suggested by the theory and reported above.

Econometric Model and Methodology

In this section I first illustrate the econometric specification of the economic model (3.1), then I briefly describe data and how to derive some of the variables that I use in the estimates (3.2). Part 3.3 presents econometric and methodological issues which came out in the study.

3.1 Econometric Model

The basic model I use for my estimations is as follows:

$$y_i = \log Y_i = \beta_1 + \beta_2 S_i + \beta_3 X_i + \beta_4 X_i^2 + \beta_5 gender_i + \beta_6 ethnic_i + u_i \qquad (4)$$

[15] If we look for example at the parents' time endowment, this is fixed by definition.

[16] See Behrman and Taubman (1986) for a criticism of this hypothesis.

[17] The authors adopt an order probit model of highest educational attainment and an OLS estimate of the logarithm of years of education (computed as the average years of schooling for each highest educational qualification). They check for non-monotonicity of the sharing rule between siblings and check the effect of the inclusion of several variables. For more details check Booth and Kee (2005).

$$\begin{aligned} S_i = \gamma_1 + \gamma_2\, age3439 + \gamma_3\, age4045 + \gamma_4\, age4650 + \gamma_5\, age5155 + \gamma_6\, mumdegree + \\ + \gamma_7\, daddegree + \gamma_8\, mumwork + \gamma_9\, mumage2125 + \gamma_{10}\, mumage2630 + \\ + \gamma_{11}\, mumage3140 + \gamma_{12}\, mumage41 + \gamma_{13}\, dadage2125 + \gamma_{14}\, dadage2630 + \\ + \gamma_{15}\, dadage3140 + \gamma_{16}\, dadage41 + \gamma_{17}\, more\_bk + \gamma_{18}\, lots\_bk + \gamma_{19}\, inner + \\ + \gamma_{20}\, town + \gamma_{21}\, village + \gamma_{22}\, rural + \gamma_{23}\, movedaround + \gamma_{23}\, familysize + \\ + \gamma_{24}\, bo\_index \end{aligned} \qquad (5)$$

In equation (4), *y* is the logarithm of labour income, *S* represents years of schooling (or, alternatively, corrected years of schooling[18]), *X* are years of experience (and appears in a quadratic functional form). Dummy variables are included to distinguish the effect of gender and ethnicity[19]. Equation (5) is used to estimate years of schooling. We are able to control for individuals age cohorts and parental education, wealth, age cohorts and attitudes to education. Dummies for areas mostly lived in as a child are included to account for environmental specificities. The base of this dummy is 'lived in the suburbs' and we account for the following alternatives: inner city, town, village, rural areas and moved around. The variable family size is the total number of siblings in respondent's family, while the variable birth order is the birth order index[20]. Furthermore, we are able to build five parental age cohorts with the base being mother and father aged less than 21 when respondent was born. Parental age cohorts may be a good proxy to estimate if younger parents have positive or negative effects on children education. If younger parents might have less experience, less financial resources and less time to spend with children because of the substitutability between working time and time spent with the family[21], it is also true that they might have more energy and enthusiasm, and focus more on the quality of the time spent with children. From the regression[22] I find not significant coefficients for father age cohorts, while the age of the mother has a relevant impact on years of schooling. In particular, the results show that mothers aged more than 26 are associated with greater schooling length for their children. This means that for our sample, more mature parents influence positively children's education.

I proxy parental family resources with a dummy variable indicating whether or not mother's respondent was working when respondent was aged 14. Family background is summarised by two dummies for parental education[23]. It is worth noting that this paper follows the approach found in Maluccio (1998) in which parents' education is an instrument for schooling. Although not reported in the tables, our measure of family background has been tested as a control both in OLS and IV regressions. I found low significant coefficients when family background is treated as a control in the wage equation and then I opted for using parental education as an instrument for years of schooling. The BHPS offers data also on the presence of books during childhood, and this information may be a good proxy for parents' attitudes to education. Furthermore, the availability of books can be a good indicator of the importance and care that parents assign to education. In fact, it is not unlikely that parents that buy a lot of books are also convinced of the usefulness of education; consequently they might put more emphasis on this aspect of children growth. Children

[18] See part 3.2 for further details on how this measure has been built.

[19] The base for the estimation is 'white male'.

[20] As pointed out in Booth and Kee (2005), although the survey is retrospective, it is unlikely that data on family size and birth order are affected by errors because it is difficult to forget the number of siblings ever had or one's own birth order.

[21] See Booth and Kee (2005).

[22] Table 4 reports the estimated coefficients of the IV regression.

[23] The use of parents' education as a proxy for family background is widely recognised in the empirical literature. See part 2.1.

might also benefit directly from the availability of books to the extent that more books represent a source of diversification of impulses and an incentive for learning.

In separate regressions I replace years of education with a corrected measures of it which takes into account educational attainment. I also test a number of different hypotheses on the functional form both of schooling and years of experience allowing for quadratic and even cubic terms[24].

In equation (4) I expect a positive sign for β_2 and β_3. β_4 will detect the magnitude of non-linearity of the experience term. 'White male' is the base for the two dummies (gender and ethnicity), so I expect a negative sign for both as frequently reported in the literature[25].

As I highlighted in part 2.2.1, while we expect a negative effect of the size of the family on years of education, the estimation of equation (5) will reveal the sign of the coefficient of the birth order index. We cannot say *a priori* that this will be positive or negative because of the simultaneous presence of forces that work in different directions.

3.2 Description of Data and Variables

The unique source of data is the British Household Panel Survey (BHPS). This is an annual survey carried out in Britain since 1991. Most of the variables have been taken from wave 13 conducted in 2003, but in order to obtain some other variables I needed to combine observations from all the waves. From the original sample (comprising more than 16 thousands observations), I select people between the age of 28 and 55[26]. Then I restrict the sample to those observations with valid information on labour income[27], years of education, birth order and family size. I also drop some observations whose labour income exceeds £100000. The final sample includes 6146 observations, 49,43% male and 50,57% female; 40,4% are first born[28], 30,4% are second born, 15,37% third born and the rest is shown in table 2. Furthermore, of our sample 20,94% are aged between 28 and 33, 24,95% between 34 and 39, 23,56% between 40 and 45, 16,19% between 46 and 50 and 14,36% between 51 and 55. The mean family size[29] is 3.34, but if we look at the mean for each age cohorts, we find evidence of the decreasing trend of the British household's fertility. From a mean of 3,67 for individuals aged 46-50 we end up with 2,88 for people aged between 28-33.

Although the BHPS does not provide direct information on years of education, it has been possible to recover this data from other variables. In particular, three questions were posed to interviewed people. In the first one they were asked about school leaving age (excluding technical college) and then they were asked if they got involved in some kind of further education and, if yes, which was the further education leaving age. I use these data to construct the variable "years of education". Table 2 shows the average years of schooling in relation to birth order. It is worth noting that the average years of education decreases when birth order increases[30]. In fact, the sample mean for years of education is 13,38 (and it is very

[24] See part 3.3.2 for more details.

[25] See among others Ashenfelter and Rouse, (1998).

[26] The choice of this range is driven from the necessity to exclude those observations for which the family size might still change and to include only people whose education is completed (see Booth and Kee, 2005).

[27] This lack of data availability is due to the fact that proxy respondent's answers have not been reported in the BHPS.

[28] Among the first born children, 7,91% are 'only' child.

[29] It is worth noting that in this study family size is the number of siblings and so does not include parents.

[30] This trend is clear up to the sixth born. There is a sort of discontinuity for seventh and ninth borns, but most probably this may depend upon the sample composition.

close to the A-level of the British Educational System), but if we look in detail this average with respect to birth order, we find that for lower birth order we have a higher average for years of schooling. From 13.64 and 13.39 respectively for first and second born, we get 13 and 12,28 for fifth and sixth born. It is also interesting noting that the average years of education for each position born is increasing in the age cohorts. Thus, as expected, this average is declining in birth order, but this trend is also affected by the family size effect.

I also use years of education to build another variable which takes into account educational attainment. This variable, labelled "corrected years of education", has been constructed in the following way. Given that in the British educational system one needs at least a certain number of years of education to achieve a qualification, when this number exceeds a threshold (for example 11 years for the GCSE or, alternatively, the O level) and is less than the number of years needed to obtain the subsequent higher qualification, the number of years of education has been reduced to that threshold. This is equal to the hypothesis that one more year of education which does not give any additional qualifications has no effect on individual earnings. In particular, I focus on three levels: the O-level (11 years of schooling), the A-level (13 years of schooling) and higher education (16 years of schooling).

As far as the individual labour income is concerned, I found data on annual labour income and not on hourly earnings[31] or hours worked. This means that the interpretation of the results must take into account this participation as well as hourly or weekly earnings.

In order to construct parental age cohorts and capture the effect of parental age at the time when respondents were born, I use respondent's year of birth and his/her parents year of birth. Subtracting the latter from the former I obtain the needed information.

A set of dummy variables has been constructed in order to account for gender, ethnicity, respondent age cohorts[32], parental family resources and other family background controls.

A more detailed description of the variables is provided in table 1.

3.3 Methodological and Empirical issues

This study is mainly focus on the use of family size and a birth order index as instruments for years of education, but a number of issues have been addressed. In the following I report the strategies adopted.

3.3.1 Why the Instrumental Variable approach?

The choice of the Instrumental Variable approach has been driven from two basic considerations. First, I intend to prove that the birth order index (built in Booth and Kee, 2005), not only has desirable features in order to isolate the effect of birth order on the amount of resources that parents devote to their children education, but also can be used to explain the choice of years of education without falling into the trap of the ability bias. Secondly, panel data techniques would not be helpful because there is very little variation over time in a given individual's education.

[31] Although a specific question asking information about hourly earnings was posed to all the interviewed, only few observations are available.

[32] The presence of a dummy for respondent age cohorts might mitigate the lack of data on institutional features of the school system.

3.3.2 Functional forms

As we saw in part 2.1, the linearity of the schooling term and the quadratic form for the experience term may be a strong assumption on the model. In order to decide which functional form to adopt in this study, I tested the following alternatives:

i. linear term both for schooling and experience;
ii. linear term for schooling and quadratic for experience;
iii. quadratic term for schooling and linear for experience;
iv. quadratic term for both;
v. cubic term for both.

Specifications i. and v. gives inconsistent estimates and are therefore excluded. In iv. I find low significant coefficients for both the experience terms and it is excluded as well. Specification iii. gives a 1% significant coefficient for the schooling quadratic term, but this is very small in magnitude. Then specification ii. seems to be the most appropriate functional form and the results reported in the tables refer to it.

3.3.3 Measurement of Income

The measurement of income is referred at a point in time (namely the occurrence of the survey) and has been dictated by the characteristics of the dataset. In fact, the BHPS provides information only on annual labour income, and it has not been possible to find enough data either on hourly earnings or hours worked per year. For this reason equation (4) explains both earnings per hour and hours worked. It means that the estimated coefficient for years of education summarize the effect of education on earnings and on the labour market participation decision.

Another issue related to the income variable regards the choice of whether or not to include individuals with zero income in the regressions. If we opt to account for zero income individuals we obtain a lower estimation because for these individuals the rate of return is null. This is misleading because these individuals might have found a well paid job after the interview and then their rate of return would be positive. On the other hand if we exclude from the sample those without income, the so called sample selection bias will work in the opposite direction. I have chosen to reduce the sample to those observations with positive income and then I expect that this source of bias will bring upward the estimated coefficients.

The choice of taking natural logarithms is justified by the fact that the distribution of log-earnings is very close to a normal distribution and Heckman and Polachek (1974) provide an accurate study on this topic.

3.3.4 Measurement of Education

In most datasets, education is affected by survey measurement error. If we assume that observed schooling may differ from actual schooling by an error term with zero mean and constant variance[33], then the OLS estimate is asymptotically biased[34]. It has been shown that most of the times the OLS estimations are affected by a downward bias that is around 10-15%[35].

3.3.5 Measurement of Experience

Since I do not have any pieces of information on years of experience I found convenient to follow a standard approach by imputing a number of years of experience equals to the individual age minus years of education minus five.

[33] Notice that the error term is uncorrelated with earnings (Card, 2001, p.1135).

[34] See Card, (1999), p. 1816.

[35] See Angriest and Krueger (1991) for a detailed description of this topic.

3.3.6 Building the birth order index

In order to explain years of education using data on family size and birth order, it is necessary to separate the effect of the two variables. The relation between them can be explained by the fact that the probability of being in a smaller family is higher for the first born child and smaller for increasing birth order. A simple correlation coefficient is in fact 0.6911 (quite high). Following Booth and Kee (2005)[36], it is possible to construct a birth order index which is able to capture the effect of birth order in a separately way from the family size effect. Let call:

$N \in [1, 10]$, the total number of siblings;

$A \in [1, 5.5]$, the average birth order[37];

$\Phi \in [1, 10]$, the respondent's absolute birth order.

The birth order index is readily defined as follows:

$$B = \Phi/A \qquad (5)$$

Checking the correlation between family size and the birth order index already defined, we find that the correlation is almost disappeared (0.0581). Moreover, we notice that, by construction, this index has an average within-family value of 1, and, more important, is constant across families. From table 1, one can see how the actual values of the birth order index means are very close to one, and in particular we get 1.006778 for women, 0.9904012 for men and 0.9986869 for the overall sample. If we put together these pieces of information we have a proof of the efficiency of this index.

3.3.7 The problem of endogeneity of education and the choice of instruments

We have shown in section 2.1 that ignoring the ability term can lead to a bias estimator of the rate of return to education. One of the main goals of this paper is to show that it is possible to deal with the correlation between years of schooling and the error term in the wage equation using two valid instruments, namely family size and birth order.

The first problem to solve is to purge the family size effect from the birth order effect, and this is done following the methodology in Booth and Kee (2005) in part 3.3.5. Even more important is to understand if these instruments meet the needed requirements in order to be valid instruments.

They must satisfy the following conditions:

i. they must be correlated with years of schooling;

ii. they need to be independent from measurement error which might affect the way we measure years of education;

iii. they must be uncorrelated with the error term in the wage equation.

Condition ii. is readily satisfied[38]. Condition i. is also satisfied and the economic literature provides a clear evidence of this fact[39]. Condition iii. is the most difficult to accept, but a number of argument are in support of this idea. First of all, if we think at ability as something strongly related to genetic factors[40], it becomes difficult to find a link between the

[36] The authors develop a study in which they explain educational attainment (and not years of education) using family size and the birth order index.

[37] For example, if N is equal to 5, A is equal to (N+1)/2=3.

[38] It is not possible to find evidence of the fact that the distribution of the birth order index and the distribution of measurement errors are somehow linked.

[39] See Becker and Lewis (1973), and Booth and Kee (2005) for a specific analysis of the influence of birth order.

[40] Studies based on twins have been already discussed in part 2.1.

birth order index and the individual ability endowment. Moreover, there are no reasons in support of the idea that (in our sample) a first born in a large family is abler than the second born in a small family[41]. Furthermore, most of the components that this index might detect are not linked with the ability factor, because the effect of this index is to define with more accuracy the resources devoted to education, and so is unlikely to be correlated with the ability factor.

3.3.8 Heterogeneity across individuals

The rate of return of education may present substantial differences across individuals. The sources of this heterogeneity can be explained through the interaction of several factors. But one of the most important ideas is that different family backgrounds can drive different outcomes in the schooling performance. But, as part 3.1 has shown, the BHPS provides enough data to control for heterogeneity across individuals.

Econometric Results

In this section I show the econometric results and give an interpretation to them. Part of the section is also devoted to the analysis of the first stage of the Instrumental Variable regression. Conclusions end the paper.

4.1 Interpretation

The interpretation of the IV estimator is one of the main challenges of this paper. In order to recognise the correct explanatory power of the coefficients, it is necessary to bear in mind all the considerations made in part 3.3.

Table 3 reports the estimated coefficients from the most significant regressions I run. Column 1 shows the coefficients for an OLS regression in which family size and the birth order index appear in the wage equation. The coefficient of the birth order index is not significant (with a p-value of 0.517), and this is in support of the choice of the birth order index as an instrument for years of education. The family size is significant at 1%, with a magnitude of -0.02, but if we compare this data with the effect of family size on years of education (as shown in table 4 column 3), it is evident that the direct effect of family size on education is much more strong. I then prefer to leave the family size variable in the schooling equation.

Columns 2 and 3 (table 3) present the simple OLS results from the original Mincer's model described in section 2.1. These regressions are the baseline for subsequent comparison with IV regressions. The OLS regression gives an estimated return to schooling of 0.058, but once I use a measure of years of education which takes into account educational attainment, I obtain a coefficient of 0.1149. The higher value of this coefficient may be interpreted in the following way. First notice that the corrected measure of schooling reduces the number of years of education by the number of years spent on education without reaching a qualification. It means that an individual with 14 years of education cannot have a degree (because he needs at least 16 years of schooling) and the years of schooling assigned to

[41] It is worth notice that a birth order index equals to 1 is assigned both to the third born in a family with five siblings and to the second born in a family with three siblings. So the probability that an individual have assigned a value for her/his birth order index is not related to unobserved ability.

her/him will be 13 (the number of years needed to get the A-level)[42]. This procedure has the effect of cutting relatively unproductive years of education and so of raising the rate of return.

Columns 4 and 5 show the OLS coefficients after including dummies for gender and ethnicity[43]. Notice that the coefficients of both measures of years of education do not vary sensitively if compared with the first two columns. The last two columns report the estimates of the Instrumental Variable regression. If we compare columns (3) and (5) we find that the IV estimator is just 2,33% above the OLS estimate. This means that most of the biases present in the estimation might offset and the result is a coefficient very close to the one found with the IV approach.

Summarizing what already said in part 3.3, in OLS regressions of the rate of return to education we can identify the following sources of bias:

Ability bias: this source normally leads to an upward bias in the OLS coefficient of years of schooling;

Measurement error: this is related to the fact that observed schooling might differ from the true value and should bring a downward bias in the OLS coefficient;

Sample selection bias: this source of bias is brought about by the fact that we excluded observations with null labour income and this should push upward the estimated coefficients.

Noticing that these sources of bias work in opposite directions, I deduce that there is a sort of balance between them. This result is on the same line of Ashenfelter and Zimmerman (1997).

The similarity between the OLS and IV coefficients raises the following question: to what extent we can proof that education and incomes are driven simultaneously by the same factors (namely ability), and, even if there are common factors, to what extent we can be sure that they exert the same impact on our variables of interest. My results show that it is possible to explain years of education with valid instruments, but since this does not alter sensitively the length of observed education, I must conclude that the endogeneity of education produces small bias for my sample and is counterbalanced by other sources of bias which push in the opposite direction.

It is also worth noting that, as brought out in Card (1999), family background variables have been used both as controls or to derive IV estimators. Differently from what Lam and Schoeni (1993) found in their study, trying to use family background as a control for variability in earnings leads to low level of significance for the estimated coefficients. I then use parental education as a proxy to determine years of education.

4.1.1 First stage IV regression

Since my estimations are performed in Stata, and for the IV regression the software shows only the results for the second stage, I reproduce the 2SLS procedure step by step and the results of the first stage are in table 4. In particular, column 1 refers to a regression which does not include any family composition variables. From column 1 to 2 we add the family size variable, and in column 3 the regression includes also the birth order index. This method allows us to check whether the estimated coefficients suffer of sensitive differences once we add family composition variables. I find that the size of the family has a negative impact on years of schooling and this effect does not vary too much once I include the birth order index[44] (the difference between them is only 0.005, in fact I obtained a coefficient equal to -0.159264 when I do not include the birth order index, and -0.1541092 if the index is used).

[42] Notice that the BHPS includes information on educational attainment and this data has been used to check the consistency of the corrected measure of years of schooling.

[43] The coefficient for ethnicity is significant at 10% only, but this might be due to the small fraction of non-white people in our sample.

[44] This result is on line to what Booth and Kee (2005) found in their study.

The analysis of table 4 shows clearly that the Booth[45] and Kee's birth order index is actually able to capture the effect of birth order separately from the family size effect.

Table 4 is also useful to understand the effect that other variables exert on years on education. In particular, I find that the effect of educated parents on years of schooling is highly significant and large in magnitude. It is also interesting to notice that the effect of father's education (1,84) is almost the double of the effect of mother's education (0.95).

We also notice the strong effect of the presence of books in parents' house during childhood. The coefficients are significant at 1% level and suggest buying more books to children!

4.2 Conclusions and possible extensions

In this study I used data from the BHPS to investigate the relationship between earnings and schooling and I found that using family size and birth order as instruments for years of education do not alter sensitively the estimation of the rate of return to schooling if compared with a standard OLS regression. This result is due to the simultaneous presence of different sources of bias which are likely to offset.

One of the key founding of this paper is that I do not find any strong evidence of the impact of the ability bias on the estimation of the number of years of education in the first stage of the IV regression. This claim, on the same line of Griliches' (1977) results, imply that we need to find out whether the ability factor affects mostly the experience term. Thus, a way to improve this study could be to find better measures for the experience term taking into account the implications of the training on the job[46].

Another possible extension might be to reduce the sample to those individuals who have at least reached the O-level and then build dummy variables for each education level. This should allow providing evidence of possible differentials in the marginal effect of subsequent qualifications that cannot emerge from this study.

[45] The authors use the birth order index in an ordered probit of educational attainment and not of years of education.

[46] See Griliches (1977). The author points out that the inclusion of a measure of training on the job might represent a way to strongly improve the fit of these kinds of models.

Tables

Table 1: Variable Means and Descriptions

Variable name	*Description*	*Women* *Obs. 3107*	*Men* *Obs. 3034*	*Total* *Obs. 6141*
age2833	Age cohort between 28-33 years old	.2088832	.2099539	.2094121
age3439	Age cohort between 34-39 years old	.2539427	.2448912	.2494708
age4045	Age cohort between 40-45 years old	.2339878	.2373105	.2356294
age4650	Age cohort between 46-50 years old	.1622144	.161503	.1618629
age5155	Age cohort between 51-55 years old	.140972	.1463415	.1436248
Gender	D=1 if respondent is female	.5059437	.4940563	
Mumdegree	D=1 if mother has degree	.0386225	.033619	.0361505
Daddegree	D=1 if father has degree	.0663019	.058998	.0626934
Mumwork	D=1 mother working when r. was 14	.5873833	.5458141	.5668458
mumage20	Mum<=20 when resp. was born	.0936595	.1123929	.1029148
mumage2125	Mum aged 21-25 when r. was born	.3089797	.2834542	.2963687
mumage2630	Mum aged 26-30 when r. was born	.28645	.2709295	.278782
mumage3140	Mum aged 31-40 when r. was born	.2481493	.2244562	.2364436
mumage41	Mum>=41 when resp. was born	.0627615	.1087673	.085491
dadage20	Dad<=20 when resp. was born	.0341165	.0421885	.0381045
dadage2125	Dad between 21-25 when r. was born	.193756	.1921556	.1929653
dadage2630	Dad between 26-30 when r. was born	.2990023	.2857614	.2924605
dadage3140	Dad between 31-40 when r. was born	.3234631	.284443	.304185
dadage41	Dad>=41 when resp. was born	.1496621	.1954515	.1722846
lots_bk	D=1 if lots of books when child	.4023173	.2824654	.3431037
more_bk	D=1 if quite a few books when child	.3533956	.3823336	.3676926
less_bk	D=1 if not many books when child	.2362407	.3269611	.2810617
Inner	Lived in inner city as a child	.0878661	.1058009	.0967269
Suburban	Lived in suburban area as a child	.2343096	.2247858	.2296043
Town	Lived in a town as a child	.292887	.2824654	.2877382
Village	Lived in a village as a child	.2088832	.2073171	.2081094
Rural	Lived in rural area as a child	.1261667	.13118	.1286435
Movedaround	Moved around as a child	.0498874	.0484509	.0491777
Ethnic	D=1 if non-white	.0231735	.0263678	.0247517
n_siblings	No. of children in r.'s family	3.372385	3.310481	3.341801
Boindex	Birth order index	1.006778	.9904012	.9986869
Yedu	Years of education	13.36402	13.40211	13.38284
Yeduc	Corrected years of education	12.5111	12.51384	12.51246
Experience	Experience	22.51786	22.56625	22.54177
exp2	Squared experience	577.1126	578.7218	577.9077
logincome	Natural logarithm of labour income	9.147139	9.750607	9.445286

Table 2: Birth order and average years of education

Birth order	*No. of obs. and %*	*Average years of education*	*Corrected average years of education*
Eldest	2481 (40,4%)	13.64289	12.74849
Second	1867 (30,4%)	13.38725	12.52062
Third	944 (15,37%)	13.1928	12.31038
Fourth	381 (6,2%)	13.13911	12.20735
Fifth	216 (3,52%)	13.00463	12.14815
Sixth	113 (1,84%)	12.28319	11.63717
Seventh	58 (0.94%)	12.31034	11.82759
Eighth	40 (0,65%)	11.925	11.425
Nineth	20 (0,33%)	12.8	12.3
Tenth or more	21 (0,34%)	11.33333	11.14286

Table 3: OLS and IV Estimates (dependent variable: logarithm of annual labour income)

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
Variables	**OLS With birth order and family size as controls**	**OLS**	**OLS***	**OLS with gender and ethnicity**	**OLS* with gender and ethnicity**	**IV**	**IV***
Years of education	.0571034 (12.68) ***	.0580759 (12.48) ***		.0576902 (12.82) ***		.0590337 (11.68) ***	
Corrected Years of ed.			.1149061 (15.73) ***		.1150325 (16.30) ***		.1097501 (12.26) ***
Experience	.025725 (2.83) ***	.0255863 (2.72) ***	.021988 (2.38) **	.0244623 (2.69) ***	.021018 (2.36) **	.0249457 (2.74) ***	.020223 (2.26) **
Exp2	-.0005153 (-2.72) ***	-.0005273 (-2.69) ***	-.0004017 (-2.07) **	-.0005094 (-2.69) ***	-.0003855 (-2.06) **	-.0005149 (-2.72) ***	-.0003802 (-2.03) **
Constant	8.82152 (63.94) ***	8.396038 (62.30) ***	7.744015 (51.45) ***	8.724253 (66.52) ***	8.063966 (55.19) ***	8.698577 (62.89) ***	8.144917 (48.27) ***
Gender	-.5991905 (-20.85) ***			-.6013953 (-20.91) ***	-.6032511 (-21.15) ***	-.6013336 (-20.91) ***	-.6032876 (-21.15) ***
Ethnic	-.1297509 (-1.40)			-.1528528 (-1.65) *	-.154781 (-1.68) *	-.1540934 (-1.66) *	-.152246 (-1.65) *
Family size	-.0277355 (-3.49) ***						
Birth order index	-.023499 (-0.65)						

OLS* and IV* refer to regressions with corrected years of schooling.
Significant at 10%; ** significant at 5%; *** significant at 1%
Number of Observations: 6146
Instruments: age3439 age4045 age4650 age5155 mumdegree daddegree mumwork mumage2125 mumage2630 mumage3140 mumage41 dadage2125 dadage2630 dadage3140 dadage41 more_bk lots_bk inner town village rural movedaround number_siblings boindex

Table 4: First stage of IV estimates with and without information on family size and birth order

	(1)		*(2)*		*(3)*	
Variables	**No family composition variables**		**Only family size included**		**Complete regression**	
Family_size			-.159264	(-6.44)***	-.1541092	(-6.23) ***
Birth order index					-.63448	(-4.95) ***
Constant	11.95995	(44.45) ***	12.53711	(44.34) ***	12.91876	(44.16) ***
Age3439	.6492509	(5.05) ***	.68433	(5.34) ***	.6743467	(5.27) ***
Age4045	.6769105	(5.16) ***	.7599381	(5.78) ***	.7211196	(5.49) ***
Age4650	.6467761	(4.45) ***	.7375076	(5.07) ***	.6621403	(4.54) ***
Age5155	.3492812	(2.30) **	.4087221	(2.70) ***	.3360766	(2.21) **
Mumdegree	1.003615	(3.95) ***	1.00408	(3.97) ***	.9483321	(3.75) ***
Daddegree	1.909224	(9.70) ***	1.875758	(9.56) ***	1.842051	(9.40) ***
Mumwork	-.0060096	(-0.07)	-.1132296	(-1.25)	-.0691443	(-0.76)
More_bk	.4638825	(4.27) ***	.3952457	(3.63) ***	.3859944	(3.55) ***
Lots_bk	1.085628	(9.51) ***	.9955336	(8.69) ***	.9793732	(8.56) ***
Inner	.0080221	(0.05)	.0655502	(0.39)	.0674706	(0.41)
Town	-.3777444	(-3.11) ***	-.343571	(-2.84) ***	-.3346829	(-2.77) ***
Village	-.626765	(-4.77) ***	-.6109485	(-4.67) ***	-.6040141	(-4.62) ***
Rural	-.3108234	(-2.05) **	-.2036924	(-1.34)	-.231818	(-1.53)
Movedaround	.4500264	(2.08) **	.4447543	(2.06) **	.4321722	(2.01) **
Parental cohorts	Yes		Yes		Yes	

Significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Correlation coefficients

	Labour income	Log labour income	Number of siblings	Birth order	Birth order index
Labour income	**1.0000**	**-**	**-**	**-**	**-**
Log labour income	**0.6935**	**1.0000**	**-**	**-**	**-**
Number of siblings	**-0.0871**	**-0.0627**	**1.0000**	**-**	**-**
Birth order	**-0.0850**	**-0.0475**	**0.6911**	**1.0000**	**-**
Birth order index	**-0.0451**	**-0.0211**	**0.0581**	**0.7115**	**1.0000**

Fig. 1a: Distribution of labour income by birth order

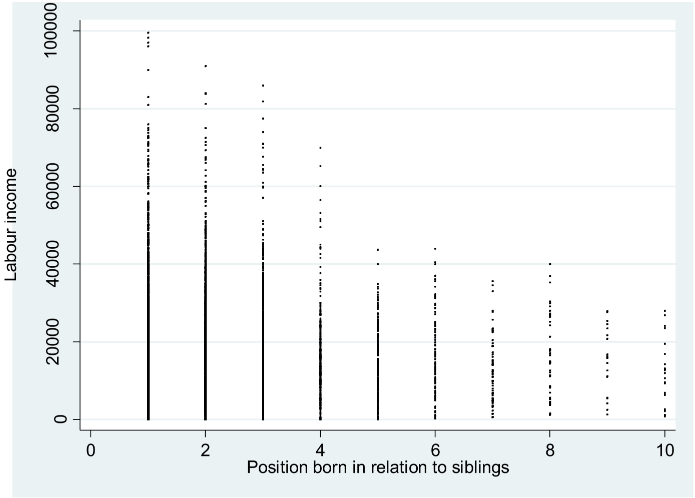


Fig. 1b: Distribution of log income by birth order

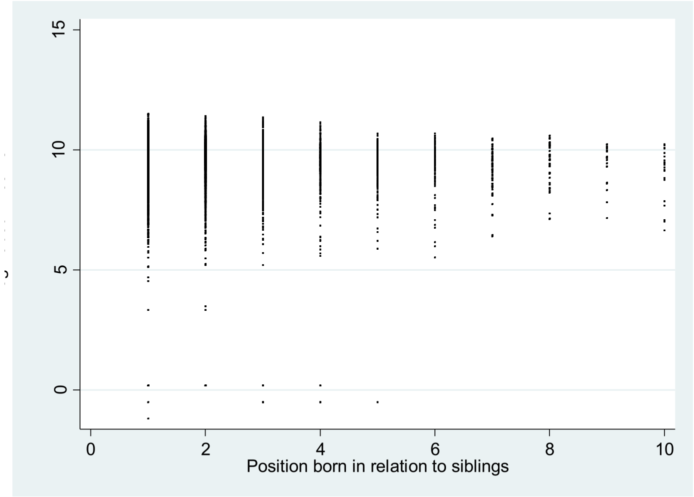
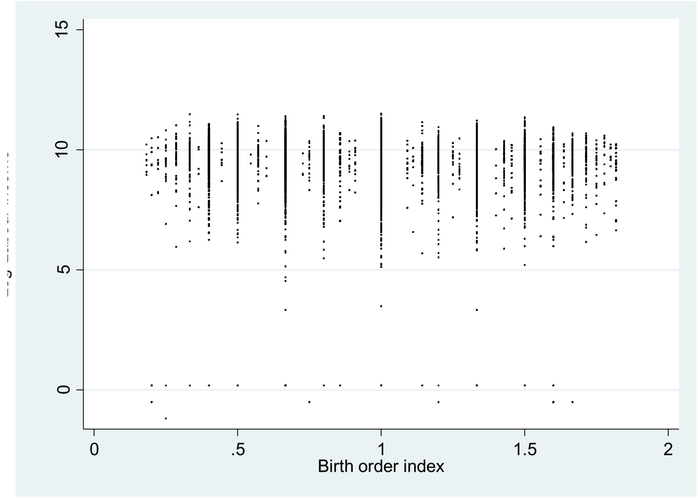


Fig. 1c: Distribution of log income by birth order index



References

Ashenfelter, O., Krueger, A.B., (1994), "*Estimates of the economic return to schooling for a new sample*", American Economic Review, Vol. 84, pp. 1157-1173.

Ashenfelter, O., Zimmerman, D., (1997), "*Estimates of the return to schooling from sibling data: fathers, sons and brothers*", Review of Economics and Statistics, Vol.79, pp. 1-9.

Angrist, J.D., Krueger, A.B., (1999), "*Empirical strategies in labor economics*", Handbook of Labor Economics, Vol. 3A, Elsevier Science.

Ashenfelter, O., Rouse, C., (1998), "*Income, schooling and ability: evidence from a new sample of identical twins*", Quarterly Journal of Economics, Vol. 113, pp. 253-284.

Behrman, J.R., Taubman, P., (1986), "*Birth order, schooling, and earnings*", Journal of Labor Economics, 4(3), Part 2, pp. S121-S145.

Becker, G.S., Lewis, H.G., (1973), "*On the interaction between the quantity and the quality of children*", Journal of Political Economy, Vol.81, pp. S279-S288.

Black, S.E., Devereux, J., Salvanes, K.G., (2005), "*The more the merrier? The effect of family size and birth order on children's education*", Quarterly Journal of Economics, 120(2).

Booth, A., Kee, H.J., (2005), "*Birth order matters: the effect of family size and birth order on educational attainment*", Centre for Economic Policy Research – Australian National University, Discussion paper n.506.

Card, D., (1995), "*Using geographic variation in college proximity to estimate the return to schooling*", in: Christofides L.N., Grant E.K., Swidinsky, R., Aspects of labour market behaviour: essays in honour of John Vanderkamp (University of Toronto), pp. 201-222.

Card, D. (1999), "*The causal effect of education on earnings*", Handbook of Labour Economics, Vol.3, Elsevier Science.

Card, D., (2001), "*Estimating the return to schooling: progress on some persistent econometric problem*", Econometrica, Vol. 69, No. 5., pp. 1127-1160.

Ejrnaes, M., Portner, C.C., (2004), "*Birth order and the intrahousehold allocation of time and education*", Review of Economics and Statistics, Vol. 86(4), pp. 1008-1019.

Griliches, Z., (1977), "*Estimating the returns to schooling: some econometric problems*", Econometrica, Vol. 45, pp. 1-22.

Hause, J.C., (1972), "*Earnings profile: ability and schooling*", Journal of Political Economy, Vol. 80, pp. S108-S138.

Heckman, J.J., Hotz, V.J., (1986), "*An investigation of the labor market earnings of Panamanian males: evaluating the sources of inequality*", Journal of Human Resources, Vol.21, pp. 507-542.

Heckman, J.J., Polachek, S., (1974), "*Empirical evidence on the functional form of the earnings-schooling relationship*", Journal of the American Statistical Association, Vol. 69, pp. 350-354.

Lam, D., Schoeni, R.F., (1993), "*Effects of family background on earnings and return to schooling: evidence from Brazil*", The Journal of Political Economy, Vol.101, No.4, pp. 710-740.

Maluccio, J., (1998), "*Endogeneity of schooling in the wage function: evidence from the rural Philippines*", FCND discussion paper No. 54, International Food Policy Research Institute, Washington.

McKinley, L.B., Neumark, D., (1995), "*Are OLS estimates of the rate of return to schooling biased downward? Another look*", The review of Economics and Statistics, No.2, pp, 217-230.

Miller, P., Mulvey, C., Martin, N., (1995), "*What do twins studies reveal about the economic returns to education? A comparison of Australian and U.S. findings*", American Economic Review, Vol. 85, pp. 586-599.

Mincer, J., (1974), "*Schooling, Experience, and Earnings*", Columbia University Press, New York.

Murphy, K.M., Welch, F., (1990), "*Empirical age-earnings profiles*", Journal of Labor Economics, Vol. 8, pp. 202 - 229.