AUTHOR ACCEPTED MANUSCRIPT

FINAL PUBLICATION INFORMATION

Returns to Education in Sri Lanka : A Pseudo-Panel Approach

The definitive version of the text was subsequently published in

Education Economics, (2015), 2015-01-30

Published by Taylor and Francis and found at http://dx.doi.org/10.1080/09645292.2015.1005575

THE FINAL PUBLISHED VERSION OF THIS ARTICLE IS AVAILABLE ON THE PUBLISHER'S PLATFORM

Title Returns to Education in Sri Lanka: A Pseudo Panel Approach

Authors Rozana Himaz (University of Oxford)

Harsha Aturupane (The World Bank)

Short Title Returns to Education in Sri Lanka

Corresponding author: Rozana Himaz (corresponding author)

University of Oxford, High Street, Oxford

e-mail: rozana.himaz@economics.ox.ac.uk or rozana.himaz@gmail.com

Tel: +44 (1865) 281800

Fax: +44 (1865) 281801

Abstract

This study employs a pseudo-panel approach to estimate the returns to education among income earners in Sri Lanka. Pseudo-panel data are constructed from nine repeated cross-sections of Sri Lanka's Labor Force Survey data from 1997-2008, for workers born during 1953–1974. The results show that for males, one extra year of education increases monthly earnings by about 5 per cent using the pseudo panel estimation rather than 9 per cent as in the OLS estimation. This indicates that not controlling for unobservables such as ability and motivation, bias the OLS estimation of returns upwards by about 4 per cent on average, driven mainly by what happens in urban areas. It also suggests that males with higher ability seem to be acquiring more years of education. This is contrary to what has been observed recently in countries such as Thailand (Warunsiri and McNown 2010) where the opportunity cost of education seems to be high, such that high ability individuals leave education for the labour market.

Keywords

Sri Lanka, Education, returns, pseudo panels, synthetic cohorts

JEL Classification codes

I00 I20 I25 C23

1. Introduction

The estimation of returns to education (both private and social) has been central to the economics of human capital since the early 1960s, and has been the subject of much debate and discussion. The most common method used to analyse private returns to education has been based on the Mincer regression of log earnings on years of schooling and years of post-school work experience (Psacharapolous 1994, Psacharapolous and Patrinos 2004, Heckman et. al. 2005). The coefficient on the schooling variable is often interpreted as an estimate of the internal rate of return. A key problem with the Mincer analysis is that it does not account for the endogeneity of the schooling variable. If unobservables such as ability and motivation are correlated with both schooling and wages, this may result in a bias in the coefficient of the schooling variable that is used to estimate private returns to education. The bias is likely to be positive if more able students are the ones who pursue more years of education. However, it could also be negative if, for example, the more able and motivated leave school early if presented with higher wage options.

The issue of unobservables such as ability and motivation causing biases in estimated returns has been the preoccupation of the literature since the earliest contributions in this area and a number of approaches have been used to deal with the issue. In some studies, measures of ability (e.g., intelligent quotient or IQ scores) are incorporated directly into the Mincerian wage equation to proxy ability. Doing so confirms that ordinary least squares (OLS) estimates are biased upwards(Blackburn and Neumark 1993). More recent non-experimental approaches have included instrumental variable methods, matching methods, and control function methods (Blundell et. al. 2004)[1]. For example Angrist and Krueger (1991), Card (1995), Harmon and Walker(1995) all attempt to instrument for schooling

[1] Studies such as Katz and Autor (1999) argue that the functional form of the Mincer model itself is no longer applicable in some cases. Their example is that of U.S. workers. This suggests that non-parametric methods may be better suited to estimate returns. Heckman, Lochner and Todd (2005) find that non-parametric estimations of returns that account for tuition costs, income taxes and non-linearities in the earning-schoolings function lead to much higher results than those based on parametric estimations of marginal rates of return for some levels of schooling in the U.S. The non-linearity of the wage-schooling relationship is not yet an issue for Sri Lanka, as basic scatter plots indicate approximate linearity. Therefore this paper will only focus on estimating returns to schooling parametrically.

outcomes with variables that are orthogonal to ability. Instruments used include quarter of birth, distance to school, the presence of a university or teacher training college in the region of residence and reforms to education policy. Often, however, the use of instruments yield returns that are higher than the OLS estimations. In some cases this is due to the instruments being weak even though they may be valid and relevant, leading to estimators that perform poorly (Bound, Jaeger and Baker 1995, Staiger and Stock 1997). In cases where the instrument used is formed on the basis of membership of some treatment group (for example, being subject to a reform in the minimum school-leaving age)[2], the higher returns reported by the IV approach may be due to the control group failing to 'neutralize cohort fixed effects' (Grenet 2013:177 Oreopoulos 2006).

Some studies have modified the more general fixed effects framework and exploited within-twins or within-siblings differences in wages and education to identify returns. This procedure yields unbiased estimates if unobserved effects are additive and common within twins (or siblings). If this is the case, that they can be differenced out by regressing the wage difference within twins against the education difference (Ashenfelter and Zimmerman 1997). A central problem with such analysis is that unobserved effects may have an individual component as well as a family component which is not independent of the schooling variable. Thus although the family component is controlled, the individual component may not be, leading to results that may not be any less biased than those of the ordinary least squares estimation. Another issue is that any error in measurement of the schooling variable will account for a larger fraction of differences between siblings than across the population as a whole. This means that forming differences between siblings will increase the bias from measurement error, causing a downward bias in estimates. One way to deal with this issue, is to adjust estimates using independent measures of error variance. Many within twin studies that control for measurement error suggest an ability bias that is relatively small (Grilliches 1979).

Another alternative is to use large cross section data sets repeated over time to create pseudo panels or synthetic cohorts as suggested by Deaton (1985), to estimate returns under certain assumptions that control for unobserved individual specific effects

[2]The impact of increased schooling is identified, in this case, by comparing average earnings between the control group and marginal group hit by the treatment. This is the local average treatment effect (LATE) interpretation of the return to an extra year of schooling. For more on the LATE and 'potential outcomes' framework see Imbens and Angrist (1994).

including those such as ability and motivation (Verbeek and Nijman 1992; Verbeek 2008). Pseudo panels are typically constructed from a time series of independent surveys conducted under the same methodology on the same reference population but in different time periods. Examples of such independent surveys are the labour force survey and household income/expenditure surveys. The pseudo-panel is created by grouping individuals into criteria that do not change from one survey to another such as the year of birth or education level of the adult, assuming there are very limited options for changing the level of formal education in adulthood.

In this paper, we attempt to estimate returns to education using pseudo panels, as an alternative to the Mincerian approach that uses cross-section data for Sri Lanka. The paper makes two important contributions to the literature. First, it is one of the few cases where pseudo panel techniques have been used to developing country data labour markets to estimate returns.. The paper, therefore, adds to a small, but growing pool of papers that attempt to re-estimate returns to education in developing countries using recent advances in the literature. Secondly, the case of Sri Lanka offers interesting insights to other developing countries. Sri Lanka continues to use education policy to attain higher welfare and lower inequality, as it emerges from a 30 year civil conflict (World Bank 2011). The use of education as a tool for development is not unusual. However, unlike many developing countries has already reached a high standard of literacy and school enrolment, which has been maintained for decades. It was hailed as an 'outlier' in the 1980s for being a country with low income but high human development indicators. So how can research be used to inform policy as to which areas in education should be targeted if indeed education should continue to be targeted? With regard to the latter point, Ganegodage and Rambaldi (2011) argue that although investment in education in Sri Lanka has contributed to growth, returns to investment in education are significantly lower than those found in other developing economies. This seems to be confirmed in Himaz (2010) *inter alia*, that estimates that returns could be as low as 4 per cent based on sibling data. Although this could be an underestimate as sibling data used comes from younger households where siblings are at an earlier stage of the earnings lifecycle, this figure is much lower than the global average return to an year of education of 10 per cent quoted in Fasih et. al (2012: 2). Estimating returns accurately is important because the case of Sri Lanka offers an interesting insight as to how returns may turn out as education levels improve but income growth is slower. In

such cases the thrust and direction of policy may have to be reconsidered. For example, if returns are indeed low and are over estimated by conventional Mincerian regressions, research may have to investigate why this occurs and at what levels of education, so that policy can be geared towards supporting necessary areas, especially if education is to be used as a key tool to improve welfare and reduce inequality.

In this paper we estimate returns using pseudo panel analysis mainly to see if the Mincerian estimation indicates a bias. The analysis is an important first step towards utilising the many good quality, rich cross section data sets available in Sri Lanka to analyse trends of a longitudinal nature using pseudo panel techniques. Under this method, a series of 9 repeated cross sections taken from the Sri Lanka Labour Force Surveys for 1997-2001, 2003-2004, 2006 and 2008, are used to construct a panel data set based on cohort means for those born between 1953 to 1974. This makes the youngest worker in the sample 24 years of age and the oldest 55 years of age. We also present estimations for returns based on cross-section data using the Mincer model for comparison. As a check of robustness and comparison, we also estimate a fixed effects model of labour returns based on within-sibling data.

The rest of the paper is organised as follows: The next section looks at the conceptual framework and empirical specification. Section 3 discussed the data. Section 4 looks at the results from pseudo panel analysis and compares it with results from OLS and sibling-based fixed effects analysis. Section 5 concludes.

2. Conceptual framework and empirical specification

The standard Mincerian earnings function (Mincer 1974) is specified as follows:

$$\mathrm{w} = \alpha + \beta_0 s + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad (1)$$

Where *w* is the natural logarithm of earnings, *s* the years of schooling and *x* is years of work experience often proxied by age. This equation can be specified for time *t* (with *t=1..T*) and individual *i* (with *i= 1....N*) as follows:

$$\mathrm{w_{it}} = \gamma + \beta_0 s_{it} + \beta_1 x_{it} + \beta_2 x_{it}^2 + \alpha_{it+}\, \varepsilon_{it} \qquad (2)$$

where *w* is the log of earnings for individual *i* at time *t*, *s* is the number of years of schooling for individual *i* at time *t* and *x* is the number of years of experience (age) for individual *i* at time *t*. The term α_{it} captures unobserved individual heterogeneity that includes ability and motivation. If this was uncorrelated with the explanatory variables s_{it} and x_{it} then the model in (2) can be estimated consistently from cross section data using OLS treating $\alpha_{it+}\ \varepsilon_{it}$ as the composite error term. However, it is likely that α_{it} is correlated to both schooling and experience (s_{it} and x_{it} respectively). If α_{it} is observable, it can be included directly into the equation. However, in the absence of such information, unobservables represented by α_{it} will cause the least squares estimation to be biased. Deaton (1985) suggests that cohorts constructed from repeated cross section data can be used to estimate a fixed effects model. The cohorts, *c,* are defined by a shared common characteristic, such that each individual is a member or only one cohort. In our case, this is the year of birth. If all observations in the cohorts are aggregated, the resulting model can be written as:

$$\overline{\mathrm{w}_{\mathrm{ct}}} = \beta_0\overline{s_{ct}} + \beta_1\overline{x_{ct}} + \beta_2\overline{x_{ct}^2} + \bar{\alpha}_{ct+}\ \bar{\varepsilon}_{ct} \qquad (3)$$

Where c=1,...C and t=1,...T and w_{ct} is the average of all monthly earnings for all individuals in cohort *c* at time *t* and similarly for the other variables in the model. If cohorts are defined by those born between 1953 and 1974, this gives us a pseudo panel of 21 cohorts over 9 time periods based on 9 cross section surveys from 1997 to 2008. Estimating β_0 from (3) can be still problematic, however, as $\bar{\alpha}_{ct}$ depends on *t*, and is likely to be correlated to x_{ct} as α_{it} was likely to be correlated to x_{it}. As $\bar{\alpha}_{ct}$ is unobservable it cannot be included directly in the estimation. However, $\bar{\alpha}_{ct}$ can be treated as a fixed unknown parameter with $\bar{\alpha}_{ct} = \alpha_c$ over time, if there existed a sufficiently large number of individual observations in each cohort (Verbeek 2008). In this case the model can be written as:

$$\overline{\mathrm{w}_{\mathrm{ct}}} = \gamma + \beta_0\overline{s_{ct}} + \beta_1\overline{x_{ct}} + \beta_2\overline{x_{ct}^2} + \alpha_{c+}\ \bar{\varepsilon}_{ct} \qquad (4)$$

As Warunsiri and McNown (2010:1618) note, all error components in (2) that are correlated with the explanatory variables have been purged from the error term in (4). This makes the fixed effects estimation consistent.

The error term in (4) can be assumed normal, independent and homoskedastic if the cohort size is fixed over time. However, if the cohorts are very different in size, this will

mean that the error term is heteroskadistic and needs to be corrected by weighting each observation with the square root of cohort size (Deaton 1985:117). As cohort size can vary in our data, we use weighted least squares estimation[3].

For the sake of comparison and to check for robustness of results, we also conduct a sibling-based fixed effects analysis on the same dataset. Under this method, unobserved heterogeneity at the household level is corrected for by using fixed effects estimation to a cluster sample, where the well-defined cluster is the household in the pooled data set. I use deviations from household means for all households where there are two or more males who are wage earners, assuming that differences are across households (if they exist). If there are unobservable fixed effects and they are significant, the constant term of the fixed effects regression would be significant and the OLS estimations would be biased. The fixed effects method addresses the issue of omitted variable bias arising from unobserved heterogeneity at the household levels only if such unmeasured attributes are common to individuals in the same household. If the unobserved household effects are random instead of being fixed, they would bias the error term and invalidate standard statistical tests. In order to test for this possibility, I estimate a random effects model using the same subsample. I then use the Hausman test to compare between the fixed and random effects estimations. The results are then compared to both the pooled OLS and pseudo panel estimates.

3. Data and descriptive statistics

The data comes from a series of 9 repeated cross sections of the Sri Lanka Labour Force Surveys for the years 1997-2001, 2003-2004, 2006 and 2008. The Sri Lanka Labour Force Survey is representative of the entire country apart from the Northern and Eastern provinces for which data was not available for most of the years. The Quarterly Labour Force Survey has been carried out since 1990 to produce estimates of employment, unemployment, labour force and basic demographic characteristics. The sampling method used is a stratified two stage sample design. In the first stage, the provinces of the country

[3] . Bernard et. al.(2011) and Sprietsma (2011) discuss the issue of heteroskedasticity in more detail including properties of the error term in terms of serial correlation, etc.

were divided into sectors. From these domains, census blocks were selected according to the population distribution (based on census data) so that the probability of selection is proportional to the size of the population in the block. In the second stage, ten households were selected from each block randomly. The entire sample was then divided randomly into four groups, one for each quarter. The sample includes all persons in the household aged 10 or above. A household or housing unit is defined as a house, apartment or room (s) that is occupied as a separate living space. The survey excludes housing units with 5 or more lodgers and institutions such as hospitals or military camps. Although the data has been collected quarterly in most years we have aggregated it by year for the purpose of this study. We use data from 1997 onwards mainly as some of the previous surveys are slightly different in design. Most notably, some of the previous surveys include only detailed information on earnings from paid employment (wages and salaries) rather than all earnings including self employment and own account work.

The nine repeated cross sections of data are used to construct a panel data set based on cohort means for those born between 1953 to 1974. This makes the youngest worker 24 years of age and the oldest 55 years of age in the sample. When the full sample is used, this generates 189 cohorts (21 years *times* 9 different surveys) to provide yearly cohort means. The average observations per cell is 318. If two year cohort means are used, 99 cohorts are generated with average cell size being 618. This is the data used for the baseline analysis. However, as males comprise nearly 80 per cent of the pooled data and as female participation in the labour market may be subject to selection bias, we construct further disaggregated panels that are based on age and gender. Although we created both age-male and age-female cohorts, we report results only for males as the female sample is likely to suffer from sample selection issues whether we use pooled individual data or pseudo panel data and therefore report a biased coefficient. Just as labour market returns maybe explained differently between males and females, requiring these groups to be considered separately, it may also be explained differently between various sectors of the economy such as rural and urban. We therefore construct panels by age and sector as well. Since urban workers are only a fifth of the entire sample, some age-urban cohorts have very few observations, even when we increase cohort size from 1or 2 years to 5 years. As discussed later, small cohort sizes (which in our case is sometimes less than 100 observations each) may result in inaccurate standard errors in pseudo panel analyses. We therefore report

results only for rural areas. We do not conduct similar panels only for females and or for those in urban areas as the cohort sizes are very small (sometimes less than 100 observations each). This is because this may result in inaccurate standard errors in pseudo panel analyses (discussed further later).

Monthly earnings used are those from the primary employment, for those who usually work for more than 35 hours a week, which accounts for over 90 per cent of the men in the sample aged 24-55 who are employed. The earnings are in real terms, deflated by the GDP deflator with the base year being the year 2002. In all 9 survey years, roughly 86 per cent of the men are employed. Of those employed, 60 per cent are wage employees, 4 per cent are employers, and around 32 per cent are own account workers. A small percentage of around 3 per cent are classified as unpaid family workers. As we use earnings rather than wages as the dependent variable, we use information for almost all employed males and females born between 1953 and 1974 in the sample.

The level of schooling an individual has obtained is reported in terms of years of schooling completed, in the surveys. This ranges from 0 to 10 with 0 referring to having had no schooling at all, 1-5 years reflecting primary schooling, 6 to 13 secondary schooling, and over 14 years of education reflecting tertiary education received from universities and polytechnics.

We also use information on some basic characteristics of the cohort such as age to capture experience and the square of age to capture any non-linearities with respect to age-related experience. In the checks for robustness, discussed later, we estimate less parsimonious specifications.

Basic descriptive statistics for the entire sample, based on individual information is presented in Table 1 below. The first column indicates summary statistics for the full sample based on pooled data that comprises 54 759 males and females born between 1953-1974. The natural log of earnings is 8.43 (about Rs. 4582 a month). The average years of education in the sample is 8.5 and the average age is 38. Around 68 per cent live in rural areas, 21 per cent live in urban areas and 12 per cent in estate areas. The second column contains information for males only, who form over 60 per cent of the full sample. The

earnings are slightly higher than for the full sample, suggesting that women earn less on average than men. The males only sample also indicates a slightly higher education.

Table 1: Descriptive statistics for those born between 1953-1974

	Full sample (males and females)	Sub sample (males)
Monthly earnings (natural log)	8.44 (0.88)	8.52 (0.11)
Education in years	8.51 (3.95)	8.62 (0.43)
Age	38.23 (6.94)	38.78 (7.02)
Aged squared	1509.50 (538.47)	1552.78 (549.6)
Urban	0.21 (0.41)	0.23 (0.06)
Estate	0.12 (0.32)	0.08 (0.04)
Individual observations	54759	29279 (=68.6% of total)

Note: Standard deviations within brackets.

4. Results

The results using the full sample are reported in Table 2 below. The first column shows the results based on OLS estimation based on individual cross section data. The estimation includes 8 dummy variables to capture yearly fixed effects. The next two columns show results based on pseudo panel data. The first of these is based on means for yearly birth cohorts from 1953 to 1974. The second is based on means for two year groups from 1953 to 1974. The analysis is conducted using two different groupings of the data to ensure results are robust. Both these regressions include cohort fixed effects.

Returns measured by OLS analysis indicates that an extra year of education increases income by around 9 per cent. The returns measured using the one and two year pseudo panel analysis is lower at around 5 per cent. The other explanatory variables included are significant as well. The age variable indicates that as age gets higher, the log of earnings are higher as well. However, as the age square variable indicates, these effects are not linear in

their increase. The OLS regression with the pooled data shows that earnings reach a maximum at age 45.1. The last column of the table reproduces results for a sibling-based fixed effects analysis on the same dataset. The data for all years is pooled and only those households that have two or more members earning who were born between 1953 and 1974 are considered. This severely restricts the sample size and only about 34 per cent of the households in the sample fall into this category. If we did not define the birth year of the individuals, around 50 per cent of the sample have two or more individuals earning. We report results only for those born between 1953 and 1974 as this is comparable to the sample we use for the pseudo-panel analysis, in spite of the sample being very restricted.

Table 2: Returns to education estimates for individual data, one year cohort means, two year cohort means and sibling fixed effects

	Individual data (cross sectional regression)	Pseudo panel One year cohort means	Pseudo panel two year cohort means	Sibling Fixed Effects
Educ	0.09***	0.05**	0.05	0.04***
	(0.00)	(0.02)	(0.03)	(0.00)
Age	0.03***	0.05***	0.05***	0.00
	(0.00)	(0.01)	(0.01)	(0.00)
age2	-0.00***	-0.00***	-0.00***	
	(0.00)	(0.00)	(0.00)	
Constant	6.66***			
	(0.09)			
Individual Observations	54759	54759	54759	18912
Cohort-year observations	-	189	99	-
Mean observations per cohort	-	318	608	-
Number of households with two or more members earning	-	-	-	8978
R-squared	0.191			
Hausman test				$\chi^2(3)=405.77$
H_0: Difference in coefficients not systematic				

Note: Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

As discussed previously, the fixed effects method addresses the issue of omitted variable bias arising from unobserved heterogeneity at the household levels only if such unmeasured attributes are common to individuals in the same household. If the unobserved household effects are random instead of being fixed, they would bias the error term and invalidate standard statistical tests. In order to test for this possibility, I estimate a random effects model using the same subsample. I then use the Hausman test to compare between the fixed and random effects estimations. The results are then compared to both the pooled OLS and pseudo panel estimates. The Hausman test, testing the hypothesis that difference in coefficients not systematic is rejected with a Chi-squared value of 26.29. Thus a fixed effects model is more appropriate. The estimations indicate a return to education that is much lower than that of the OLS or pseudo panel analysis. This result, however, should be interpreted with much caution as only 34 per cent of the sample have households with more than one member working who is also born between the years 1973 and 1953.

The results in Table 2 seem to suggest that using cross section data and pseudo panel data yield results that are very different, with the OLS estimation overestimating returns by about 4 per cent. However, traditionally, returns are estimated separately for males and females. This is because labour force participation rates of males and females are quite different, as are the earnings and other labour market related characteristics. This means pooling males and females together and including a gender dummy to capture gender effects such that only intercept effects are captured is not sufficient[4]. Returns to

[4] The specification in Table 2 was changed to include a gender dummy as well as a gender dummy interacted with education as a part of unreported robustness checks on the reported results. When gender is included in the OLS estimations, it is a dichotomous variable. These dichotomous variables, if included in pseudo panel analysis will appear as proportions. For example, the proportion of males out of those born in 1953 in the 1998 LFS survey was 67% while in it was 72% in 2006. Thus although the pseudo panel analysis specification can include variables such as gender and sector of residence, its interpretation would not be identical to that of the corresponding OLS coefficient. The male dummy in an OLS regression would reflect how much extra a male earned compared to a female. However, the proportion of males in a pseudo panel specification will reflect how much extra would be earned if the proportion of males was increased by one unit. In our unreported results, the modified OLS specification indicates that being male males exerts a significant positive impact on wages. The results also show that the gender variable interacted with the education variables is significantly negative, indicating that the extra earnings being male due to an extra year of education is actually negative, albeit by a very small number. Put differently, for males, the impact of an extra year of education on extra wages is 0.088-0.0004. For females it is 0.088. The pseudo panel estimations do not indicate the proportion of males in the cohort as being a significant determinant of average return to education. Thus increasing the proportion of males in the sample does not increase returns statistically significantly. The interacted term is not significant either.

education could be explained differently between males and females such that not just the intercept term but the slopes (i.e., estimated coefficients on the explanatory variables) could be different as well. We therefore disaggregate and perform the analysis of returns only for males and females separately. However, as explained in section 3, we report only results for males as it is well known that estimations for females should correct for sample selection biases. This has been discussed much in the context of the conventional Mincerian specification (Heckman 1976) but not in the context of the pseudo panel analysis. This means that even though we can attempt to correct for the issue when using OLS analysis, it is less clear how this can be done in the context of pseudo panel data. Moreover females comprise only about 30 per cent of the sample, which leaves cell sizes too small for efficient standard errors is the pseudo panel analysis.

Table 3 reports results for a sub sample based on males . Returns measured using OLS analysis in column 1 indicates that on average an extra year of education increases income by 9 per cent. Quite notably, the returns measured using pseudo panel techniques reported in columns 2 and 3, are lower at around 5 per cent. This suggests that when unobservables such as ability and motivation biases upwards the returns to education in cross section analysis. In other words, males with higher ability seem to acquire more education in the Sri Lankan context. This result is in sharp contrast to what Warunsiri and McNown (2010) find for Thailand, where the returns measured using pseudo panel data are higher than that measured using OLS estimation suggesting that the opportunity cost of education is high for high ability workers. Their results suggest that higher ability workers leave formal education sooner than lower ability workers due to perhaps attractive wage rates in the labour market.

Our results also show that experience is an important contributor to earnings, with every year of experience adding to roughly 4 per cent increase in earnings in both the cross-section and pseudo panel results.

It has been observed previously that returns to education are higher in urban areas compared to rural areas (Aturupane 1993, Himaz and Aturupane 2011). Therefore the analysis was refined further by the construction of synthetic cohorts for age-male-rural

cohorts. As columns 4, 5 and 6 show, the average returns is quite close under both methods with OLS estimation indicating 8 per cent and pseudo panel methods indicating 7 per cent. This implies that the vast difference in returns between the two methods previously noted must be largely driven by what happens in urban areas. To elaborate, ignoring ability and motivation significantly over estimates returns to average education in urban areas. Unfortunately this could not be verified using pseudo panel analysis as although we constructed synthetic cohorts for the age-male-urban category, individual cell observations fell below 100 in some cases , even when 5 year cohort means were used.

Table 3 Males: Returns to education estimates for individual data, one year cohort means and two year cohort means

	Individual data (cross sectional regression)	Pseudo panel One year cohort means	Pseudo panel two year cohort means	Individual data cross sectional regression (rural)	Pseudo Panel one year cohort means (rural)	Pseudo Panel two year cohort means (rural)
	(1)	(2)	(3)	(4)	(5)	(6)
Education (years)	0.09***	0.05**	0.04	0.08***	0.07***	0.07*
	(0.00)	(0.02)	(0.03)	(0.00)	(0.03)	(0.04)
Age	0.04***	0.05***	0.05***	0.03***	0.05***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
age2	-0.00***	-0.00**	-0.00**	-0.00***	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	6.94***			7.08***		
	(0.11)			(0.13)		
Observations	37,574	189	99	29,279	189	99
R-squared	0.15			0.14		

Note: Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

5. Conclusions

This paper is a first attempt at utilising a vast wealth of good quality repeated cross section data in Sri Lanka to build pseudo panel data sets for analysing complex socio-economic issues. Using pseudo-panel or synthetic cohort analysis helps to control for unobservables such as ability or motivation that may otherwise bias the returns to education estimated by a conventional Mincerian earnings equation. The paper looked at returns to education in Sri Lanka based on conventional cross section data using the Mincerian equation, and compared it with pseudo panel data constructed using 9 Labour Force Surveys from 1997 to 2008. When men and women are pooled together, estimated returns indicate that OLS results overestimate returns by nearly 4 per cent compared to using pseudo panel analysis. We then disaggregate men and women as returns maybe explained differently among the two groups. We report results for males and as for the pooled analysis it is clear that the OLS analysis has a notable ability bias that inflates the coefficient on the schooling variable: An extra year of education increases monthly earnings by about 9 per cent using the OLS estimation rather than 5 per cent as in the pseudo panel estimation. This is suggestive that not controlling for unobservables such as ability and motivation bias the OLS estimation of returns upwards by about 4 per cent on average for males. Further refining the cohorts by not just age and gender but sector as well, shows that the over estimation of returns in Mincerian specification seems to be driven by what happens in urban areas. This suggests that in Sri Lanka males with higher ability, particularly in urban areas, seem to be acquiring more years of education contrary to what has been observed recently in countries such as Thailand (Warunsiri and McNown 2010) where the opportunity cost of education seems to be high such that the more able seem to leave formal education to join the labour market. Unfortunately, we are not able to perform the same analysis for women and obtain reliable results as the cohort sizes become too small.

These results are just a first step and there is much more that needs to be done, in terms of analysing urban returns, returns to females and returns within categories of education before concrete policy recommendations can be made. Overall, however, the basic results are indicative of the fact that the endogeneity of schooling does seem to inflate returns estimated using OLS analysis as is consistent with similar work done for countries such as the US. The implication of this result that labour market incentives (and

opportunity costs) may not be high enough for the more able students to abandon formal education is encouraging from a policy perspective. It suggests that further investment in quality and quantity of education can support either the productivity enhancing or screening roles of education.

References

Angrist J. D and A.B. Krueger, Does compulsory school attendance affect schooling and earnings?, *Quarterly Journal of Economics* 106 (4) (1991), pp. 979–1014.

Ashenfelter. O and D, J. Zimmerman, "Estimates Of The Returns To Schooling From Sibling Data: Fathers, Sons, And Brothers, *The Review of Economics and Statistics,* MIT Press, vol. 79(1) (1997), pages 1-9, February.

Bernard, J, D. Bolduc and N Yameogo, A pseudo-panel data model of household electricity demand, *Resource and Energy Economics* 33 (2011) p 315-325.

Blackburn, M L and D Neumark. "Omitted-Ability Bias and the Increase in the Return to Schooling," *Journal of Labor Economics*, University of Chicago Press, vol. 11(3) (1993), pages 521-44

Bound, J and D. Jaeger and R. Baker, Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association* 90 (430) (1995), pp. 443–450.

Blundell, R, L. Dearden and B. Sianesi, Estimating the returns to education: Models, methods and results, Centre for the Economics of Education, London School of Economics, London (2001).

Card, D. Using geographic variation in college proximity to estimate the return to schooling. In: L . Christofides, K. Grant and R. Swindinsky, Editors, *Aspects of labour market behaviour: Essays in honour of John Vanderkamp*, University of Toronto Press, Toronto (1995).

Deaton, A. Panel data from a time series of cross-sections, *Journal of Econometrics* 30 (1985), pp. 109–126.

Fasih, T, G. Kingdon, H.A.Patrinos, C. Sakellariou and M. Soderbom, Heterogeneous Returns to Education in the Labor Market, *World Bank Policy Research Working Papers* (2012) http://dx.doi.org/10.1596/1813-9450-6170

Ganegodage, K. R and A. N. Rambaldi, The Impact of Education Investment on Sri Lankan Economic Growth. *Economics of Education Review* 30 (2011), 1491– 1502

Grenet, J. Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws, *Scandinavian Journal of Economics*,

vol. 115, no. 1 (2013), pp. 176-210,.

Grilliches , Z Sibling Models and Data in Economics: Beginnings of a Survey, *Journal of Political Economy,* Vol 87:5. S37:S64 (1979)

Harmon, C & Walker, I,"Estimates of the Economic Return to Schooling for the United Kingdom," *American Economic Review*, vol. 85(5) (1995), pages 1278-86.

Heckman, J. J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," NBER Chapters, in: Annals of Economic and Social Measurement, Volume 5, number 4, pages 475-492 National Bureau of Economic Research, Inc (1976).

Heckman, J., Lochner, L., and P Todd. . *Earnings functions, rate of return, and treatment effects: The mincer equation and beyond.* NBER Working Paper: 11544 (2005).

Himaz, R 2010. Intrahousehold Allocation of Education Expenditure: The Case of Sri Lanka, *Economic Development and Cultural Change*, University of Chicago Press, vol. 58(2), pages 231-258, 01.

Guido W. Imbens, G and J. D. Angrist Indentification and Estimation of Local Average Treatment Effects, *Econometrica* Vol. 62, No. 2 (1994), pp. 467-475

Katz, L. F. and D. H. Autor. Changes in the Wage Structure and Earnings Inequality, in O Ashenfelter and D. Card, *Handbook of labor economics*. Vol. 3A. Amsterdam: Elsevier Science, North-Holland(1999), pp. 1463–1555.

Mincer, J Schooling, experience, and earnings, National Bureau of Economic Research, New York (1974).

G. Psacharopoulos, Returns to investment in education: A global update, *World Development* 22 (9) (1994), pp. 1325–1343.

G. Psacharopoulos and H.A. Patrinos, Returns to investment in education: A further update, *Education Economics* 12 (2) (2004), pp. 111–134.

Oreopoulos, P. The Compelling Effects of Compulsory Schooling: Evidence from Canada. *Canadian Journal of Economics*, 39(1) (2006), pp. 22-52.

Sprietsma M. Computer as pedagogical tools in Brazil: a Pseudo-panel analysis, Education Economics, 20(1) (2012), 19-32.

Staiger D and James H. Stock, "Instrumental Variables Regression with Weak Instruments," Econometrica, Econometric Society, vol. 65(3) (1997), pages 557-586

M. Verbeek and T. Nijman, Can cohort data be treated as genuine panel data?, *Empirical Economics* 17 (1992), pp. 9–23.

Verbeek, M. 2008 Pseudo Panel s and Repeated Cross Sections in L. Matyas and P.Sevestre, The Econometrics of Panel Data: Fundamental and Recent Development in Theory and Practice third edition. Springer.

Warunsiri S. and R. McNown. The return to Education in Thailand: A Pseudo-Panel Approach, World Development 38(1) (2010) 1616-1625.

World Bank (2011). Transforming school education in Sri Lanka: From cut stones to polished jewels. Colombo: World Bank.