CONVEXITIES AND THE RETURNS TO EDUCATION IN SOUTH AFRICA

by

Laura May Poswell

A dissertation submitted to the School of Economics at the University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Business Science.

Supervisor: Malcom Keswell
Cape Town, 2003

Abstract

This paper explores the rate of return to education in South Africa with special focus on the pattern of returns to different levels of schooling. Although a basic assumption of neoclassical human capital theory is that returns to education diminish, past analysis for South Africa suggests that returns are more likely to be convex. Estimates, however, are based on widely differing data and estimation methods so that only tentative conclusions can be drawn. In this paper we undertake a rigorous econometric exercise in which the same parametric and semi-parametric techniques are applied to a number of nationally representative datasets so that the results can be effectively compared. We find that in South Africa private returns to education increase, as appears to be the case for many other African countries. The implications of such a core assumption of neoclassical human capital theory being violated are particularly relevant when analysing poverty and inequality. The neoclassical conclusion of convergence of the income distribution over time does not hold. Rather, the emergence of poverty traps and persistence of inequality becomes a likely result.

TABLE OF CONTENTS

List of Tables

List of Figures

1. Introduction

Human capital theory is based on the notion that people expend time and money in acquiring education in order to increase their productivity and consequently the expected value of lifetime earnings (Becker (1964); Hansen (1963); Mincer (1958), (1974); Ben-Porath (1967)). The distribution of investment in such human assets provides the basis for the distribution of income and wealth in an economy. Education is therefore seen as a key determinant of welfare with predictions of neoclassical human capital theories providing the basis for much development policy around the world.

In the last 20 years a key aspect of this international development strategy has been the focus on the importance of primary as opposed to tertiary education. This approach has been based on often dated, highly aggregated empirical findings of diminishing marginal returns to education. South Africa too has followed this development trajectory with government allocating significant proportions of the national budget to expenditure on schooling, particularly primary. It is therefore especially important to have knowledge of the monetary returns individuals receive when attaining different levels of education. The main task of this study is to establish the pattern of returns to education in South Africa and to consider the implications this has for the persistence of inequality.

There is evidence to suggest that returns to education in South Africa, as well as in many other African countries, may be convex as opposed to concave. Estimates however are based on widely differing techniques and data sources so that only tenuous conclusions can be drawn. We undertake a rigorous econometric exercise in which the same estimation methods are applied to a number of nationally representative South African datasets in order to establish a robust result. Only once this is achieved are we in a position to apply the appropriate economic theory when analyzing education and its impact on the earnings distribution, the implications for the persistence of inequality and the relevant policy issues that should be considered.

In this study we assume that human capital is the key determinant of income and that schooling equates to human capital. *Schooling* refers to primary, secondary and tertiary education. The paper proceeds as follows. Section 2 explores the notions of intergenerational mobility and poverty traps and explores the consequences of non-linearities in the returns to investments in human capital. In Section 3 the concept of a private rate of return to education is defined and the Mincerian earnings function derived. Sources of bias involving the Mincerian approach are also discussed. Section 4 reviews evidence of the worldwide pattern of returns to education and considers in detail recent studies of South Africa. In Section 5 we present our own empirical analysis in which we apply the same cross-sectional econometric techniques to 4 different datasets in an attempt to determine whether consistent patterns emerge. We also employ semi-parametric techniques to uncover any patterns that parametric approaches may conceal. Finally, we conclude.

2. Human Capital and Inequality: the consequence of non-linearity

Neoclassical economic theory assumes that markets function perfectly and human capital investments are subject to diminishing marginal returns. Based on these conditions, wage rates and capital-labour ratios converge within and between countries to a single, steady-state equilibrium that is stable over time. Initial endowments of income and assets do not matter for long run levels of consumption and output (Romer (1986)). In a country such as South Africa, with one of the highest levels of inequality in the world, the theory then predicts that in the long run, less-equipped economic agents (or countries) can accumulate the necessary human capital so that differences are eliminated. But what if markets do not function perfectly or the returns to investment in human capital do not decrease at the margin? Indeed, it is becoming increasingly common for empirical work to reveal that returns should not be assumed to be concave, whether these be returns to physical capital, social capital or to education. What are the consequences for poverty and inequality if this fundamental assumption of neoclassical theory does not hold? In order to gain insight into these types of issues it becomes necessary to have an understanding of the broader theory of intergenerational mobility and poverty traps.

Poverty traps occur when identical agents become 'caught' in a position of high-level or low-level economic status, due to self-enforcing initial conditions that matter not only in the short run, but in the long run as well. In such scenarios, inequality can be shown to persist for long periods. There is a range of models in which the interaction of human capital and income provide sufficient conditions to produce such multiple competitive equilibrium outcomes that are stable over time. In these models, poverty traps are shown to result from mechanisms such as credit market imperfections, externalities associated with "membership groups" and through increasing returns to human capital.

One of the most commonly modelled causes of a poverty trap is the existence of credit market imperfections. Loury (1981) and Barham, Broadway, Marchand and Pestieu (1994) construct overlapping generations models in which returns to education decrease but a liquidity constraint exists. It is not possible to borrow for training purposes from the capital market and therefore education must be financed from family income. In Loury's model, parents are altruistic and in each period family income is divided between consumption and training of offspring. The parents' utility is a function of current consumption and the future well-being of their offspring, which is in part a function of the amount of training the child receives. An individual's income is also dependent on randomly distributed innate ability. In the model of Barham *et al*, parents are not altruistic, but a child can borrow from his/her parents, provided he repays the loan in the second period. While he is being educated the loan is used to finance consumption and the cost of training. In both these models, the extent of investment in the child's education is determined by the initial wealth of the parent.

In Loury's model, if parents investing less in children who experience higher marginal returns to education could borrow from parents investing more in children who receive lower marginal returns, then income in the next period could on average be higher for both families. In imperfectly working credit markets, this convergence does not take place. In Barham *et al* where income is divided between consumption and training, and a child must repay money borrowed to finance training, there may be a level in which family income is so low, that it is rational for an individual not to acquire an education at all. He will choose rather to work and earn in period one.

The key point to note is that in both of these two models the existence of multiple equilibria is possible. Intergenerational mobility is low and those with initially low earnings may be trapped with low human capital and in poverty.

Another set of models of persistent inequality consider the division of society into particular social groups as an important determinant of poverty traps. Lundberg and Startz (1996) explore the role of human capital accumulation in generating persistent inequality among white and black Americans. They develop an overlapping generations model in which positive externalities associated with a stock of social capital impact on future individual human capital investments which in turn impact on the stock of social capital accumulated. In times of racial discrimination, the pattern of investment in human capital would reflect the expected return to education for each of the races. If the expected return is lower for African Americans than for equally educated white Americans the optimal strategy for Africans would be to invest in a lower level equilibrium. Lundberg and Startz (1996) show that if social capital is a function of the average human capital of a group, and if social capital exhibits increasing returns, then inequality can persist over time. This need not be due to present discrimination, but rather might be due to past discrimination that altered the pattern of human capital accumulation in an earlier period. Clearly this is a relevant finding for South Africa.

Along similar lines Durlauf (2002) explores how segregated and different groups impact on members in different ways. His example of "role model effects" explores a situation in which the probability of going to college depends on the percentage attendance of one's "group". He shows how multiple steady states of college attendance may arise in the absence of current discrimination but rather as a result of group attitudes and perceptions.

We finally consider a third set of models which are the main interest of our paper. In these models that once again implicitly place human capital as the key determinant of economic outcomes, it is the actual pattern of returns to investment in human capital that impact on the persistence of inequality. In a world in which the returns to human capital increase, as opposed to decrease, the implications for the possibility of poverty trap outcomes can be severe.

Romer's (1986) seminal work on increasing returns and growth shows the possibility of long run divergence in the distribution of income between countries in the presence of increasing returns to human capital investment. His research was motivated by the evident persistence of inequality between countries at different stages of development. Romer develops a competitive equilibrium endogenous growth model in which knowledge is the basic form of capital, the accumulation of which by forward-looking, profit maximizing agents drives long run growth. The model comprises three key elements relating to the production and accumulation of "new knowledge", and its impact on output growth. The first assumption regards "new knowledge". "New knowledge" results from research that is subject to diminishing marginal returns in the accumulation of further knowledge. When at the frontier of research and development, it is especially difficult for a firm to acquire more knowledge. "New knowledge", however, gives rise to positive externalities in that if one firm acquires new knowledge, the production possibilities of other firms expand as they can copy or learn from the innovating firm. Production of consumption goods is a function of, among other things, the stock of knowledge, which exhibits increasing returns in the production of output. Even if all other factors of production are held constant, it will not be optimal to stop acquiring knowledge and reach a state in which no further research is undertaken. Production in the second period is a function of consumption and knowledge acquired in period one. Romer shows that in a two period model in which consumption must be foregone so that new knowledge can be produced, multiple steady states can arise and persist in the long run in response to small shocks or initial differences. In this model diminishing returns to accumulation of knowledge ensure that production does not overheat and a positive externality is necessary for equilibrium to exist. It is increasing returns, however, that allow multiple equilibria to obtain[1].

Considering the pattern of returns from the opposite stance, Tamura (1991) develops an endogenous growth model that predicts convergence of income growth rates and per-capita income levels. The motivation for this exercise was to explain the convergence of living standards between developed countries. In his model, income converges because human capital accumulation (also seen as knowledge) is subject to diminishing marginal returns. It is easier for those with below average knowledge to acquire knowledge from the existing knowledge pool, than those with above average knowledge to acquire "new knowledge". His assumptions differ to those of Romer's in that the human capital spillover is in the production of human capital, not in the production of consumption goods. The model predicts that where there are heterogenous agents, differentiated only by their initial human capital, convergence will occur and result in a homogenous population.

In the two models above, it is the concavity or convexity of returns to human capital that is responsible for the single or multiple equilibrium outcomes. In these models human capital enters as a continuous variable. The following two models build on the work of Romer (1986) and Lucas (1988) but introduce the increasing returns framework through a discontinuous or otherwise termed "non-convex" relationship between human capital and income.

Galor and Zeira (1993) model a situation in which the existence of both imperfect credit markets and non-convex education can generate a poverty trap. In the first period of their two period overlapping generations model individuals have the option to invest in education or work as unskilled labour. In the second period, those who invested in education in period one work as skilled labour and those who did not invest in education continue to work as unskilled labour. In

[1] See Lucas (1988) for a similar model.

period two individuals work, consume and leave bequests. An individual's utility is a function of consumption and bequests in period two. Individuals are identical except for differences in their initial wealth. The return to skilled labour is greater than the return to unskilled labour so that individuals would prefer to obtain training. In their model it is more expensive to borrow than to lend so that it is easier for those with initial wealth to invest in education. Furthermore, the acquisition of human capital is indivisible. One can only invest in a fixed amount of human capital in period one. It is this non-convexity which gives rise to a strong form of increasing returns to education. If one does not possess sufficient wealth for the "minimum investment" necessary to acquire education, and if the cost of borrowing is too expensive, one will remain unskilled in the future. Those with higher wealth will be able to invest in education, reap the higher returns and will remain wealthy. Rich and poor dynasties result and the inherited distribution of wealth persists.

Bardahn and Udry (1999) also assume liquidity constraints and non-convex education. Once again, an individual is either educated or uneducated. Individuals cannot borrow to finance their studies. Rather they must forego consumption, save and then pay for education out of their savings. In a steady state, there is a ratio of skilled to unskilled workers that generate a return to education that equals the return to physical capital. For any skilled to unskilled labour ratio that is less than that of the steady state, the return to education will be greater than the return to physical capital and skilled labour will invest in education of future generations. If uneducated workers have no assets, there is a possible steady state in which they rationally choose not to invest in education of future generations and rather to remain unskilled. This is because an individual has to finance education out of savings. To do this, he must forego consumption. If the ratio of skilled to unskilled is very low, the wages of the unskilled are very low and the utility of current consumption becomes very high. There then exists a steady state where it is too expensive for the poor to become educated. The cost of foregoing consumption for long enough to save sufficiently to acquire education, becomes too high.

In the above examples, if the returns to investing in human capital diminished and markets functioned perfectly as neoclassical theory predicts, then people could slowly accumulate education and convergence in human capital and the income distribution would result. If this core assumption of neoclassical economics is violated, however, and returns to education are shown to increase, there is a real possibility that low level and high level equilibria may occur simultaneously. In such circumstances the focus of policy must be to 'break the divide'. Small investments in many individuals may in fact not be useful. The investment required to pull someone out of poverty might rather be extremely large. Although very little work of the type described above has been done for South Africa (the few examples include Posel (1999), Keswell (2001) and Hertz (2001)), it becomes apparent that in order to understand more fully the debate about inequality, we should consider the various factors that may lead to low intergenerational mobility within a poverty trap framework. Either on their own or in combination credit market imperfections, social capital spillover effects and increasing returns to human capital can be used to explain why multiple steady states arise and why inequality might persist in the long run. The implications for policy are then quite different to those in an analysis where markets function smoothly and returns to human capital diminish, with equality resulting in the long run.

3. Defining Returns to Education: the Mincer Model

In this section we develop the Mincerian model used to estimate the rate of return measure on which we focus in this paper. We also consider potential difficulties with the measure so that we can be aware of its limitations from the outset.

There are typically two methods in the human capital literature that are used to measure the pecuniary returns to education. The first and more complete method is to employ traditional cost-benefit analysis. When measuring returns to education according to this method, one considers the present value of future income received from additional education net of the earnings foregone and other investments made in acquiring the relevant education level. The *internal rate of return* on the investment is the discount rate that equates *the present value of costs incurred and the opportunity cost of earnings foregone with the present value of earnings arising from the additional education acquired.* The second method approximates the internal rate of return through an earnings function estimation approach. This method has data requirements that are often easier to fulfil than the full cost benefit analysis approach, and is applied through regression models utilising cross sectional data. Below is a brief presentation of Mincer's choice theoretic framework of this model (Mincer (1974:10-11)).

Using the assumption that the time span of a person's earnings life is fixed, let

n = length of working life
Y_s = annual earnings of an individual with s years of schooling
V_s = present value of an individual's lifetime earnings at the start of schooling

r = discount rate

t = 0, 1, 2,..., n time, in years

d = difference in the amount of schooling in years

e = base of natural logarithms

Then for a discrete discounting process the present value of lifetime earnings at the start of schooling is

$$V_s = Y_s \sum_{t=s+1}^{n+s} \left(\frac{1}{1+r}\right)^t \tag{1}$$

and for a continuous discounting process

$$V_s = Y_s \int_s^{n+s} e^{-rt} dt = \frac{Y_s}{r} e^{-rs}(1-e^{-rn}) \tag{2}$$

and

$$V_{s-d} = Y_s \int_{s-d}^{n+s-d} e^{-rt} dt = \frac{Y_{s-d}}{r} e^{-rs}(1-e^{-rn})e^{-r(s-d)} \tag{3}$$

We let $V_s = V_{s-d}$ in order to find the ratio $k_{s,s-d}$ of annual earnings after s years of schooling to annual earnings after s-d years of schooling. It is in this way that we find the marginal rate of return or discount rate that equates the present value of earnings for the two different levels of schooling

$$k_{s,s-d} = \frac{Y_s}{Y_{s-d}} = \frac{e^{-r(s-d)}}{e^{-rs}} = e^{rd} \tag{4}$$

Now we can define

$$k_{s,0} = \frac{Y_s}{Y_0} = k_s \tag{5}$$

and from the equation above we have $k_s = e^{rs}$ (6)

Taking logarithms we get

$$\ln Y_s = \ln Y_0 + rS \tag{7}$$

which is the canonical earnings function attributed to Mincer.

The constant term ($\ln Y_0$) accounts for expected earnings in the absence of any education. The coefficient on years of education is interpreted as the marginal internal rate of return on education where the costs of education are accounted for in terms of foregone earnings attributed to the time spent in attaining a certain level of schooling (Rosen (1992)). Without including the costs of education incurred by government, the coefficient then approximates what is known as the ***private rate of return*** to time spent in school.

Experience enters the equation as a quadratic polynomial to more fully account for the importance of on-the-job learning in the concave age-earnings profile. The relationship between schooling and earnings is linear (both in S and r) in this specification. The general form of the equation is as follows with S referring to the years of schooling, X to experience and u to the stochastic error term.

$$\ln Y_s = \ln Y_0 + rS + b_1 X + b_2 X^2 + u \tag{8}$$

Assessing how the returns to education may differ by education level, can be incorporated into the model in various ways. Using dummy variables for each year of education allows a unique rate of return for each level of schooling. An alternative is to incorporate higher order schooling terms into the earnings function. In such an equation there will also be a different rate of return for each schooling level.

Mincer's earnings function has another useful application. It can be used to measure the earnings inequality that is attributable to differences in educational attainment. Simply taking the variances of the basic earnings equation gives a measure of the distribution of earnings as a linear function of the distribution of schooling. This is a proxy for earnings inequality attributable to schooling inequality. Taking variances of equation (7) we get

$$\sigma^2_{\ln Y_s} = r^2\sigma^2_S + \sigma^2_u \tag{9}$$

(Mincer 1974: 25). The coefficient of determination (R^2) indicates what fraction of the variation in schooling explains the variation in log earnings. Focusing on the r^2 term, it can be clearly seen from equation (9) that as the rate of return to education increases, so too does earnings inequality.

Issues of bias concerning measurement of Mincerian returns to education

The Mincerian return is only an approximate measure and it is necessary to be aware of the number of ways the estimate may be biased, either in an upward or downward direction. The main areas of bias that have been the subject of much research consider the role of ability, family background and school quality as well as measurement error that arises from the misreporting of educational attainment.

Omission of ability from the earnings function is a potential cause of bias in OLS returns to education estimates. If there is a positive correlation between education and ability, so that those who have greater ability tend to stay in schooling longer, then earnings that are attributed to higher levels of schooling may rather be a function of higher ability. When ability is unobserved and omitted from earnings function calculations, the estimate of the return to education will be biased upwards. Attempts to quantify the ability bias have used instrumental variable approaches and data using twins to control for ability differences. Some studies attempt to control for "observable" ability using test scores such as IQ tests as measures of cognitive ability. What is understood by ability itself is a contentious issue in its own right and whether test scores accurately capture ability is another area of debate. Empirical evidence on the magnitude of ability bias (when ability is omitted from the earnings function) tends to suggest that it is relatively small. Griliches and Mason (1972) find this affect to be around 12 percent while Griliches (1977) finds it to be markedly lower and not necessarily in an upward direction. In a survey of studies on investments in education, Schultz (1988) places the figure as most likely falling between 5 and 15 percent.

Omitting family background or social status indicators from the earnings function may also result in biased coefficients, either due to the potential correlation between genetics and ability or through the possibility that wealthier and "better socially connected" parents will secure more education for their children, as well as high-paying jobs (Schultz (1988)). In a comprehensive review of recent studies of returns to education Card (1999) finds that parental or sibling education is most likely to have a small positive impact on earnings.

A positive correlation between quality and quantity of education and the omission of a quality variable in the earnings function may also result in education estimates being biased. Measuring school quality in terms of the relationship between test scores and wage rates has generally indicated a limited association between school quality accounted for in this way and earnings (Schultz (1988:590); Card and Krueger (1992:1)). Research using other indicators of school quality such as teacher education levels, teacher: pupil ratios and relative teacher salaries points to a much more convincing and significant relationship between school quality and earnings. For example, in a study of Brazilian males aged 15 to 35, Behrman and Birdsall (1983) use average teacher education in the area an individual acquired his schooling as a proxy for schooling quality. On comparing OLS estimates of standard earnings functions with estimates of specifications including the quality variable they find that the "omission of quality" bias on the traditional estimate of the private return to education is a very large 75 percent in the upward direction. Using a two stage fixed effects model on United States data, Card and Krueger (1992) find that men schooled in states with higher quality education systems (measured by relative teacher pay, average term length and the pupil-teacher ratio), earn higher returns to their educational investments. Case and Yogo (1999) find corroborative evidence for black South African males where the magisterial district in which one is schooled, has a large and significant effect on the rate of return to schooling. The extent to which omission of quality from OLS regressions biases estimates in the latter two studies, is not however quantified. It is important when interpreting returns to education, to be aware that the simple rate of return estimate to quantity includes the relationship between omitted quality of schooling and its correlation with educational attainment (Schultz (1988)).

Whereas omission of ability, family background and school quality are most likely to bias OLS estimates in an upward direction, measurement error resulting from individuals misreporting their educational attainment will in most cases bias results downwards. Using data on identical twins Ashenfelter and Krueger (1994) find that omission of ability variables does not bias returns to education estimates upwards but that measurement error in reported education levels biases results downwards significantly. Recent work by Kane, Rouse and Staiger (1999) cautions about the implications of measurement error when the measurement error is non-classical. This might occur as those with the lowest level of schooling cannot underreport and those at the highest level cannot overreport their educational attainment. They find that those with completed college education have a higher probability of reporting schooling levels correctly than those with less than completed college and that OLS estimates of the returns to education will tend to be understated for incompleted college and overstated for college completion. Card (1999) finds that the downward bias in conventional Mincerian schooling coefficients due to measurement error is probably in the order of 10%. When family background effects are controlled for, the bias is more likely to be in the region of 15%.

The final type of bias considered here may be particularly important for a country like South Africa. This is sample selection bias that is introduced when using ordinary least squares to estimate relationships on a censored sample. In a country such as South Africa with its exceptionally high rate of unemployment, all representative surveys contain a high proportion of unemployed or zero earners. Through excluding zero earners and by not controlling for the probability of finding employment, ordinary least squares estimations may produce biased results. Appendix C details why such biases might arise. Furthermore, if the total effect of education is the subject of interest and if we are to accurately capture the impact that education has on the individual, not only in terms of raising one's earnings once employed, but also in terms of how education fares in helping one to find employment in the first place, all labour market participants should be included in our analysis. It is for this reason that many recent studies using South African data have opted for techniques such as the tobit or Heckman that are better suited to censored samples (see for example Bhorat and Leibbrandt (2001); Keswell (2001); Rospabe (2001)).

The extent that the various forms of bias matter in terms of the final coefficient on the education term depends on the nature of data used and relatedly on the country for which the analysis is being performed. Using data for the USA Griliches (1977) finds that when allowing for schooling measurement errors as well as omitted variable bias, the biases appear to offset each other. A similar result is found with UK data by Deardon (1999). She finds that the effects of measurement error bias and what she terms "composition bias"[2] on the OLS schooling coefficient

[2] Composition bias refers to the differences that may occur between individuals who self select into employment and is equivalent to selection bias discussed above. If the characteristics of those who have jobs and those without differ, then another potential source of bias exists.

almost directly offset the impact of omitted ability and family background bias. Such findings may hold true for the developed world but it should be borne in mind that issues of bias may have a greater impact on results generated from data for developing and less egalitarian societies. Using 1993 data on black South African males Hertz (2001:15) in fact finds that the magnitude of the biases is high in the South African context. Hertz (2001) assigns 1.2 years of education for every completed grade of schooling to control for the1.2 years on average it takes for the African males in this sample to complete a level at school. The time spent in graduating from one standard to the next is therefore greater than one year and the corresponding earnings foregone relate to this longer period. Hertz (2001) controls for omitted variables through household fixed effects. Such fixed effects models, however, tend to exacerbate problems of measurement error. Using panel data available from a second wave of the survey carried out in 1998, he develops an estimator to control for measurement error by comparing the responses on schooling of those who self-reported educational attainment in both waves of the survey with those whose schooling was reported by another household member in either wave[3]. Upon correcting for both omitted variable bias and measurement error he finds that conventional Mincerian return estimates are upwardly biased in the region of 50%.

In summary then, the Mincerian rate of return to schooling, albeit an approximate measure, is simple to obtain through the application of regression analysis to cross-sectional data. It is therefore a commonly used measure of the return to education. Even though there are a number of ways in which the estimate may be biased, it is not clear that the biases will necessarily change the pattern of the returns to different levels schooling, which is the subject of this paper. Nevertheless, the potential biases should be taken into account when specifying earnings functions and interpreting results.

4. Existing Evidence: Global and Local

In Section 2 a number of models were described in which the pattern of returns to education are shown to play a fundamental role in explaining why inequality might persist in the long run. For theoretical analyses involving the study of intergenerational mobility and poverty traps, the importance of correctly ascertaining whether returns are increasing or decreasing becomes

[3] In further estimations of returns to education Hertz (2001) uses higher order polynomials with the relationship between schooling and earnings non-linear. In order to control for bias in these more complicated specifications he estimates by both fixed effects (which results in estimates being biased downwards) and ordinary least squares (which leads to estimates being biased upwards) and takes his preferred result as the average of the two.

evident. In Section 4.1 we consider international evidence from the last 40 years and how this has influenced the notion that returns to education diminish. In Section 4.2 we review existing evidence on South Africa in an attempt to establish if consistent patterns with regards to the returns to schooling emerge.

4.1 International Evidence

Early work on returns to education (see Becker (1964); Hanoch (1967); Hansen (1963) and Mincer (1974)) suggested that returns were most likely to be diminishing with the greatest returns to education accruing to primary levels of schooling, followed by secondary and then tertiary. Such findings have shaped much of the thinking around the relationship between earnings and education. They have led to the generally accepted supposition that returns are concave and formed the building blocks of theories of growth in which diminishing returns is a core assumption and a sufficient condition for the attainment of single steady state equilibria. The implication for the distribution of income is that given time, convergence will occur.

The most comprehensive work done in collating and comparing international estimates are the cross-county analyses by Psacharopoulos (1973, 1985, 1994 and Psacharopoulos & Patrinos (2002)). These studies carry through the notion that returns diminish by level of education. "The classic pattern of falling returns to education by level of economic development and level of education are maintained" (Psacharopoulos (2002:1)) Closer inspection of the actual figures, however, highlights some important caveats.

Table 1 below shows the Mincerian return by region with mean years of schooling given where available.

Table 1: The Coefficient on Years of Schooling: Mincerian Rates of Return to Education

Region	2002 Mean per capita ($US)	2002 Mean Years of Schooling	2002 Coefficient	1994 Mean Years of Schooling	1994 Coefficient	1985 Coefficient
Sub-Saharan Africa	974	7.3	11.7	5.9	13.4	13
Latin America/ Caribbean	3 125	8.2	12.0	7.9	12.4	14
Asia*	5 182	8.4	9.9	8.4	9.6	11
Europe/ Middle East/ North Africa*	6 299	8.8	7.1	8.5	8.2	8
OECD	24 582	9.0	7.5	10.9	6.8	9
World	9 160	8.3	9.7	8.4	10.1	11

Source: Psacharopoulos & Patrinos (2002), Psacharopoulos (1994), Psacharopoulos (1985)
* Non-OECD
Note: The figures for each region comprise simple unweighted averages compiled from the estimates for all countries falling into that region.
The specified regions do not represent the same group of countries in each paper. As more country-specific studies have become available, the dataset has increased so that the 2002 analysis includes many more countries per region than the 1985 and 1994 papers.

It appears that only in the 1994 dataset, do the coefficients decrease consistently by level of schooling. The pattern is similar for 1985 but the latest summary statistics for 2002 are not as precise. Specifically, Sub-Saharan Africa has, on average, smaller returns than Latin America, even though the average educational attainment is almost one year lower in Africa. Furthermore, the OECD countries average both higher mean years of schooling and a higher rate of return than the non-OECD countries in Europe, the Middle East and North Africa. The most recent results do not necessarily imply diminishing returns.

The cross-country studies also compare both *social* and *private rates of return* by level of education for each region. Inclusion of all costs incurred by society (for example school subsidies and teachers salaries) leads to a measure known as the *social rate of return to education*. This measure is helpful in guiding government investment decisions with regards to funding different types of schooling. Our study here is limited to the private rate, which according to human capital theory should impact on incentives and investment decisions of the individual. It is however, useful to be familiar with the pattern of social rates of returns and how this has impacted on general thinking about the relationship between educational attainment and the return to education. Table 2 incorporates Psacharopoulos's findings for the three latest studies.

Table 2: Average Social Returns to education levels measured as percentages calculated according to the "full method"[4]

Region	2002 Primary	2002 Secondary	2002 Tertiary	1994 Primary	1994 Secondary	1994 Tertiary	1985 Primary	1985 Secondary	1985 Tertiary
Sub-Saharan Africa	25.4	18.4	11.3	24.3	18.2	11.2	26	17	13
Latin America/ Caribbean	17.4	12.9	12.3	17.9	12.8	12.3	26	18	16
Asia*	16.2	11.1	11.0	19.9	13.3	11.7	27	15	13
Europe/ Middle East/ North Africa*	15.6	9.7	9.9	15.5	11.2	10.6	13	10	8
OECD	8.5	9.4	8.5	14.4	10.2	8.7	na	11	9
World	18.9	13.1	10.8	18.4	13.1	10.9	23[a]	14	12

Source: Psacharopoulos & Patrinos (2002), Psacharopoulos (1994), Psacharopoulos (1985)
* Non-OECD
[a] average excluding OECD

It can be seen from the above that on average social returns do appear to diminish, in the studies from 1985 and 1994. It is such findings that have led to the persistence of the belief that social returns diminish by education level, so that from a public policy perspective primary schooling is found to be the most desirable investment and tertiary the least. This pattern holds for the lesser-developed regions in the 2002 study but not for Europe/ Middle East/ North Africa and for the OECD. The latest study is interesting as it brings to light new evidence of how the pattern of social returns is likely to change as countries on the development frontier grow and advance further. However, again the assertion that returns diminish by level of education is not incontrovertible.

Finally, we consider the evidence on private returns by education level within region. Even though the returns are private these studies are not directly comparable with Mincerian returns as they are calculated according to the *extended cost benefit* method. The rates are also average rates to a level of education relative to no education, rather than marginal rates from one level to another[4]. Nevertheless, it is the pattern that is of interest to us and the average rates should reflect a similar pattern to that calculated when using the Mincerian approach extended to allow for different levels of education (using dummy variables or education splines).

[4] The average rate of return is calculated according to the extended method using the following formula

$$\sum_{t=-S}^{0} (C + W)_t (1 + r)^{-t} = \sum_{t=1}^{n} (W_S - W_0)_t (1 + r)^{-t}$$

where S = years of schooling, C = the costs incurred in acquiring S years of schooling and W = the earnings foregone in this period. W_S = the wages earned by someone with S years of schooling and W_0 = the wages earned by someone with 0 years of schooling. r is the internal rate of return to S years of schooling that

Table 3: Average Private Returns to education levels measured as percentages calculated according to the "full method"

	2002			1994			1985		
Region	Primary	Secondary	Tertiary	Primary	Secondary	Tertiary	Primary	Secondary	Tertiary
Sub-Saharan Africa	37.6	24.6	27.8	41.3	26.6	27.8	45	26	32
Latin America/ Caribbean	26.6	17.0	19.5	26.2	16.8	19.7	32	23	23
Asia*	20.0	15.8	18.2	39.0	18.9	19.9	31	15	18
Europe/ Middle East/ North Africa*	13.8	13.6	18.8	17.4	15.9	21.7	17	13	13
OECD	13.4	11.3	11.6	21.7	12.4	12.3	na	12	12
World	26.6	17.0	19.0	29.1	18.1	20.3	31 [a]	18	20

Source: Psacharopoulos & Patrinos (2002), Psacharopoulos (1994), Psacharopoulos (1985)
* Non-OECD
[a] average excluding OECD

The studies of average private returns reveal a different picture to those of social returns. In every case, private returns to primary education are higher than returns to secondary education. However, for the most part returns to higher education are found to be larger than those to completed secondary education. In fact, it is only the OECD countries in the 1994 study that exhibit on average a diminishing private rate of return to education. As the 2002 study has only recently been released (September 2002), it appears that it is probably the previous evidence of diminishing returns to both social and private education in the OECD countries that has been carried through and become the most cited and generally accepted result. As a large proportion of economic modelling is undertaken in these countries it is unsurprising that the assumption of diminishing returns is central to many theories and models. We must, however, be alert to the fact any such theories do not necessarily apply in the case in which returns to education do not diminish.

It is also clear that there is a time element involved in which the Mincerian returns appear to have decreased slightly over time corresponding to an increase in average educational attainment. Psacharopoulos & Patrinos (2002) find the same feature when analysing a number of average returns estimates from different time periods for specific countries. The implication is that the increase in supply of schooling has led to a small decrease in returns. Using the full cost benefit analysis method Psacharopoulos (1994) finds that private returns have decreased for both primary and secondary education but increased for higher education. This is an interesting result even though it does not necessarily carry through to the more recent data.

equates the present value of the costs of S year of schooling with the present value of the benefits for a working life of n years. (Adapted from Psacharopoulos (1973))

The finding is explained in Carnoy (1995) in which changes in rates of returns over time are studied for the United States of America, Columbia, Hong Kong, Kenya and Korea. Carnoy (1995) finds that in periods of rapid industrialization, combined with increased access to primary and secondary education, the rates of return to schooling appear to decline for each level over time.[5] Specifically, returns to primary fall first, followed by secondary and lastly tertiary. He finds that not only do the rates decline, but to the extent that the pattern of diminishing returns is changed to one of increasing returns. Tertiary returns end up higher than secondary and secondary returns higher than primary. This is an important finding as it highlights the difficulties of aggregating across countries and also explains why it might be expected that returns to education may be convex as opposed to concave.

Indeed, closer inspection of Psacharopoulos's results shows that many findings are extremely dated and many countries are excluded from the analysis altogether. Consideration of a number of more recent studies of African countries (all of which employ earnings functions approaches) reveals some interesting patterns. Using data from 1994 and 1995 Siphambe (2000) finds that in Botswana, Mincerian returns to primary schooling are the smallest, followed by lower secondary, tertiary, and then upper secondary. Like South Africa, Botswana has high unemployment (21%), especially among the youth, and high income inequality with a gini-coefficient of 0.52 in 1994. Skyt Nielsen and Westergård-Nielsen (1998) find a similar result for Zambia using 1993 data in which Mincerian returns to post-primary education exceed those to primary for urban men and woman[6]. Teal (2001) finds evidence of strongly increasing private returns to education in a study of Ghana using data as recent as 1999. Evidence for Egypt mirrors these other studies with Wahba (2000) finding increasing returns to education using data from 1988. Appleton, Hoddinott and Mackinnon (1996) present further estimates of increasing Mincerian returns for Côte d'Ivoire, Kenya and Tanzania. It appears that in Africa (as in the rest of the world), one should have no a priori expectation that the returns to education diminish. The possibility of convex returns is becoming increasingly likely and must be analysed on a country-by-country basis if one is to apply appropriate economic models and policy prescriptions. We now focus our analysis on South Africa and review studies that have been conducted using South African cross sectional datasets dating from 1990.

---

[5] This holds for both social and private returns for some of the countries studied and depends on the data available.
[6] There are insufficient observations to analyse data on tertiary education or to get significant results for the return to post-primary education for rural dwellers.

4.2 South African evidence

In recent years several nationally representative household surveys have become available in South Africa resulting in a burgeoning volume of econometric work focusing on labour market issues. Cross-sectional earnings functions have been widely used to derive estimates of factors such as returns to education. Policy debates over labour supply matters are heavily informed by the empirical evidence from such statistical analyses. The results, however, are difficult to compare owing to different definitional conventions, estimation techniques and sampling frames. In terms of the foundations of both theory and policy, it appears that it would be an extremely valuable exercise to establish a reliable picture of the pattern of returns to education in the South African context. Only once this has been determined, can we begin to analyse the high levels of inequality from an appropriate standpoint.

In this part of the analysis, a number of South African studies that have incorporated some measure of the rate of return to education are reviewed. Whereas the focus of some of the studies was explicitly schooling in South Africa, for others the estimates considered are one part of a wider analysis. We now attempt to collate and compare the results in order to understand comprehensively what past research has revealed and whether consistent patterns emerge.

Making cross study comparisons is however complicated by a number factors: First, one is considering 9 datasets and any difference in sampling, questionnaire design and data collection will more than likely influence results even when the surveys aim to be nationally representative.

Second, the sample under consideration often differs. Specifically, some researchers run separate regressions for each race, others for gender and still others for union as opposed to non-union members or those living in rural or non-rural areas. The age specification often differs with some authors only including those between the ages of 15-65 while others restrict or expand the age range.

Third, the dependent variable itself ranges from log of hourly wage, to log of monthly wage to actual wages and covers a range of full time, part time, formal, informal, casual and temporary workers.

Fourth, the choice of whether to include the unemployed is crucial in terms of the outcomes one is wishing to measure. If one is only interested in the contribution of education to explaining variation in earnings for the employed, then wage earners will be the focus of the study. It has been shown, however, that in South Africa not just earnings, but also the probability of finding employment is strongly associated with education (Keswell (2001); Rospabe (2001); Hofmeyer (2001)). If the *total effect* of education is the subject of interest, then all economically active individuals should be included.

Fifth, estimation approaches vary greatly from study to study. We are confronted with an array of regression types from simple ordinary least squares, to censored regression models and methods for multiple outcomes such as multinomial logits. We then also find that the specifications differ, with the number and type of controls included in estimations varying widely.

Sixth, as the focus is education, it is imperative to take account of the different forms in which education enters the models and the subsequent interpretations. In the simplest Mincerian equation, education enters linearly into the function as "years of education". Studies on South Africa have used either the actual years of education variable, or attempted to take account of non-linearities in returns by using splines for different levels of schooling (see for example (Mwabu and Schultz (2000); Bhorat and Leibbrandt (2001); Rospabe (2001)), dummy variables for each year (see Moll (1996), Lam (1999); Hofmeyer (2001); Hosking (2001)) or through higher order education terms (see Kingdon and Knight (1999); Erichsen and Wakeford (2001); Hertz (2001)). When for example, education and age enter the earnings function additively, the implicit assumption is that the differential effect of education is constant for all ages. If however, the education effect differs *with* age, then this effect can be included in the earnings function in the form of interaction terms (see Hertz (2001)).

In Appendix A we report on studies that have incorporated some form of education control and have drawn conclusions regarding the returns to schooling from their results. The table includes the relevant survey data used, the sub-sample of individuals considered, the controls included in the regressions, the type of regression applied and therefore whether sample selection has been controlled for or not, the form of the dependent variable and most importantly the terms for education and their related coefficients. Where there are many alternative specifications presented in a paper, those that are most relevant to this paper in terms of ease of comparability across studies have been included.

It can be seen from the table that the primary difficulty in trying to relate one study to another is the array of functional forms that the schooling variables take in the various regressions. Understanding how to correctly interpret these education coefficients is crucial if we are to draw any substantive conclusions from this analysis. We therefore provide a brief explanation of the different functional forms.

The education spline has been used repeatedly in South African examples in which it most often takes the form of dividing the returns to education into three categories, namely 7 years (primary), 12 years (secondary) and 15+ years (tertiary). The coefficient on the spline measures the average rate of return to an additional year of the given level of education (for example, primary, secondary and tertiary). When the dependent variable is in natural logarithm form, it is necessary to take the antilog of the spline coefficient and subtract 1 to arrive at the most accurate rate of return for the specification (see Halvorsen and Palmquist (1980)).

Using dummy variables for each year of education allows for a changing rate of return from year to year. Incremental returns are calculated by subtracting the coefficient of a given year from the coefficient of the previous or following year. The average annual rate of return from, for example, 8 years to 10 years is calculated by subtracting the coefficient on the 8 year dummy from that on the 10 year dummy and then dividing by the difference, in this case 2. It is important to take cognisance of the fact that even if the coefficients on dummy variables are higher with each year of education, this does not mean that the marginal returns are increasing. It is necessary to find the incremental return to each year for the pattern of marginal returns to be revealed. Once again, when the dependent variable is in log form, one must take the antilog of the relevant return (be it incremental or average) and subtract 1.[7]

A third manner in which to bring education into the equation is to allow for non-linearities in returns through introducing higher order polynomials in education. The coefficients in such a specification will not only denote the size of the return but the respective signs will indicate whether the returns are increasing or decreasing. In such a specification there is a different return

[7] Taking the antilog of the coefficient on a spline or dummy is a procedure that although correct, is often not performed, as the effect it has on the estimated return is usually negligible. This is indeed the case when coefficients are small and for most estimates studied here. When coefficients are large, however, the difference can be notable, as can be seen in Table 4 and in Appendix A in which Hosking's (2001) results change dramatically when the estimates are treated in the correct fashion.

associated with every year of education. The marginal rate of return is calculated by taking the derivative with respect to schooling and then calculating the return for each level of schooling.

From the above we see that it should be possible to manipulate estimates across functional forms in such a way that they can be made roughly comparable. Table 4 presents an abridged version of the table in Appendix A, where such a process has been attempted. The basis for comparison is the primary-secondary-tertiary spline, as this is the most aggregated measure and cannot be broken down any further. The returns reported should therefore be interpreted as the average private rate of return to an additional year of either primary, secondary or tertiary education.

Further difficulties arise in terms of different model specifications and most obviously different sample subjects. Running separate regressions for those of different race, gender, union status or location is likely to produce differing results to having a regression which includes all sample subjects and then utilises dummy variables to control for these sub-groups. In Table 4 below we have attempted to achieve overall comparability by taking *weighted* averages of results to attain a general figure for the population. (Weights used reflect the percentage of the relevant subgroup represented in the sample).

Table 4: Comparison of the average rate of return to an additional year of education at the primary, secondary and tertiary level across South African studies

	Survey	Sample considered	Returns to education[c] 7 yrs	12 yrs	15 yrs
Originally Splines					
Mwabu and Schultz (2000)	PSLSD, 1993	African (male)	0.09	0.17	0.34
		All races and genders*	0.05	0.21	0.37
Rospabe (2001)	OHS, 1999	All races and genders*	0.03	0.10	0.18
Bhorat (2000)	OHS, 1995	Skilled workers (male and female, Africans & whites)*	0.02	0.06	0.23
		Semiskilled workers (male and female, Africans & whites)*	0.01	0.14	0.00
Michaud and Vencatachellum (2001)	PSLSD, 1993	Africans (male and female)*	0.03	0.10	0.10
		African (male)	0.05	0.12	0.12
Bhorat, et al (2001)	OHS, 1995	Africans (male and female)*	0.04	0.10	0.03
		African (male)	0.04	0.11	0.04
Originally Dummies[a]					
Lam (1999)	OHS, 1995	All races (male)	0.09	0.25	[0.21-0.31][b]
Moll (1996)	CSS/HSRC, 1990	African (male)	0.02	0.11	[0.16-0.17][b]
Hofmeyer (2001)	OHS, 1999	African (male, formal sector union & non-union)*	0.02	0.09	[0.26-0.32][b]
Hosking (2001)	Population Census, 1996	All races and genders	0.07	0.74	[0.26-0.38][b]
Originally Higher Order[a]					
Kingdon and Knight (1999)	PSLSD, 1993	All races and genders	0.04	0.09	0.13
Erichsen and Wakeford (2001)	PSLSD, 1993	All races and genders*	0.04	0.12	0.17
		All races (male)	0.05	0.14	0.19
	OHS, 1995	All races and genders*	0.05	0.12	0.17
		All races (male)	0.05	0.14	0.19
median return			0.04	0.12	0.19

Notes:
A * indicates that where separate regressions for population sub-groups were reported, a weighted average across sub-groups has been taken. The average reflects the weighting of the subgroup relative to the sample population based on the sample sizes used for the respective regressions.
[a] For dummy variable studies, the marginal return to each level of education is calculated by subtracting the dummy for one level from that of the previous level. In the higher order studies the marginal effect to each year of education is calculated by taking the derivative with respect to schooling and then calculating the return for each year of schooling. For both type of manipulations the marginal returns are then averaged for primary, secondary and tertiary levels to calculate the 'mock spline'.
[b] There is a difficulty in calculating the marginal return to tertiary education when using dummy variables. This is because tertiary education is not homogenous and represents a variety of post secondary educational qualifications that take differing amounts of time to complete. It is therefore necessary to make an assumption about the number of years spent in acquiring a diploma or degree. The bracketed values indicate a range with the first figure calculated according to the assumption of 3 years spent in higher education and the second figure calculated according to the assumption of 2 years.
[c] Where applicable the antilog-1 of the relevant coefficient has been taken.

Upon examining Table 4, certain patterns become evident. For every study, the returns to an additional year of primary education are low and below 10%. The median across studies is in fact a mere 4%. Many authors have commented on reasons for the relatively low returns to primary education in South Africa. Mwabu and Schultz (1996) hypothesise on the excess supply of African workers with primary education and the downward pressure this would place on the wage. Moll (1996) postulates that poor schooling quality for Africans is likely to be the major cause of low returns. Historically, African primary schools had high teacher-pupil ratios relative to those of other races and a high percentage of unqualified teachers. Yet another reason why returns to primary education may be low in South Africa relative to those in other developing countries involves the structure of employment in the economy and the specific basis for large benefits accruing to primary level schooling in other developing regions. Colclough (1982) explains that primary education is particularly useful in raising the productivity of agricultural peasants. This has occurred predominantly in countries where farming is modernizing and becoming more technologically advanced. Primary schooling equips farmers with the necessary skills to adopt new production methods such as new seed varieties or marketing channels. In a review of 18 studies of 13 developing countries and controlling for area, land, capital and labour time, Lockheed *et al*, (1980) (in Colclough (1982)) find that in a modernizing environment, 4 years of primary schooling is associated with a 10 percent annual increase in a peasant farmer's output relative to that of a peasant farmer with no education. As South Africa does not have a comparable peasant-farming culture, such potential gains from primary education will not be realized.

The returns to an additional year of *secondary education* range more widely across studies. The median value is 12%. An extremely clear pattern, however, emerges in that the returns to secondary education are greater than those to primary. This is common across all studies and differs from Psacharopoulos's (1985, 1994, & Psacharopoulos & Patrinos (2001)) general worldwide pattern of primary returns being well in excess of secondary returns (see Table 3).

The returns to tertiary education fluctuate considerably and are in certain instances more difficult to interpret. This is largely due to the differences in the type of tertiary education available and the survey questionnaire design.[8] Moreover, the returns to different types of higher education

[8] In order to calculate the marginal return to a year of schooling, the time spent in acquiring that level of schooling must be incorporated. Dummy variables for a particular level or type of tertiary education often include different time dimensions thus rendering interpretations at the margin exceptionally crude.

will vary greatly. We find a median return to an additional year of tertiary education of 19%. Abstracting from dummy variable studies (for the reasons discussed above) it is evident that there are two studies that do not show that the returns to tertiary education necessarily being higher than the returns to secondary[9].

Considering more closely Michaud and Vencatachellum's analysis (2001), it becomes clear that when running separate regressions for union and non-union members, the returns for both male and female union-members increase by level of education. This pattern does not hold for non-union members. There are two probable reasons for this. Firstly, the sample of non-unionised African males with tertiary education is particularly small and secondly the regressions are run including a significant 'skilled-employment' dummy. This dummy is likely to be highly correlated with tertiary education and is in all likelihood diminishing part of its partial effect.

The other study of interest in this regard is Bhorat and Leibbrandt (2001). The authors use what they term a "three stage sample selection model" for which coefficients on the education splines are reported for each stage. They estimate equations for labour market participation, employment and earnings sequentially using probits in the first two stages. In the employment probit, the coefficients are as expected. The spline for tertiary education is significant in influencing the probability of having a job and is much higher than for secondary and primary returns. In the earnings equation, however, the coefficients on tertiary education are insignificant. The authors comment that this infers that although tertiary education is important in raising the probability of employment, it is not relevant as a predictor of the level of earnings for Africans. This result is in direct contrast with both Hofmeyer (2001) and Rospabe (2001) that find tertiary education as a significant estimator in both their employment and earnings equations. The inconsistency could reflect differences in the datasets used (Hofmeyer (2001) and Rospabe (2001) use the OHS, 1999), or it could be a consequence of inaccurate application of the econometric techniques. Specifically, Bhorat and Leibbrandt (2001) report on using the Heckman two-step approach when estimating the inverse Mills ratio in stage one for its inclusion in stage 2. The Heckman estimator, however, provides incorrect estimates of the standard errors of the coefficients. (Breen (1996:17)) No mention is made of correcting the variance-covariance matrix for such errors from stage one to stage two which could lead to biased results. Furthermore, as with Michaud and

[9] The term on the tertiary spline for Bhorat's (2000) semi-skilled workers is insignificant. We ignore this result as semi-skilled workers are unlikely to have tertiary education and so the estimate is not relevant to this discussion.

Vecatachellum (2001), a high association between the tertiary spline and dummy variables for professionals and perhaps managers is probably eroding the explanatory power of the tertiary education coefficient.

Although not exhaustive, in the majority of cases presented in Table 4, the average return to an additional year of tertiary education exceeds that of an additional year of secondary education. This result corresponds to the evidence for Non-OECD countries (see Table 3). It is also unsurprising for a country like South Africa in which global technological advances and growth of microelectronics has meant that much of the employment that has been created in the last decade has been in skilled professions (Bhorat (2001a)). Between 1990 and 1998, formal sector employment of unskilled and semi-skilled workers decreased by 19 percent whereas employment of highly skilled labour rose by 12% (Edwards (2002)). During the apartheid era, access to higher education was severely limited for all race groups other than Whites. This has resulted in the economy being skills constrained, even though it is experiencing mass unemployment of unskilled and semi-skilled labourers[10]. In light of the skills shortage, a significant premium to tertiary education would therefore be expected.

From the above analysis it appears that secondary returns are greater than primary returns and that tertiary returns are probably greater than secondary returns. Such an increasing pattern across the three education levels is indicative of a convex relationship between education and earnings in South Africa. The relationship seems plausible for a country that has taken a growth path experienced by South Africa in the last decade. Increasing access to primary and secondary education will have led to increased supply of workers with these levels of education and in all likelihood would have depressed returns. The surplus labour economy and high unemployment rate will serve to aggravate this situation. Furthermore, increasing demand for skilled labour will lead to a skills premium that will increase the return to higher education relative to lower levels. The apparent increasing returns appear highly plausible and indeed seem to reflect a similar pattern to other African countries discussed previously (see section 3.1).

Many of the South African studies reviewed do not only present estimates of the pattern of returns to education, but also make comparisons across different sub-groups of the population (for example, by race, gender or location) in cases in which separate regressions have been run. Conclusions are then drawn regarding both within-group and between-group inequality.

In the South African studies considered in this section, almost all regressions run by race, gender, union status and location exhibit the pattern of within group increasing returns. This in turn implies that we might expect to see increasing within group inequality. In fact in an analysis of discrimination, Moll (2000) finds that from 1980 to 1993, within race earnings inequality has increased. Convexity in the returns to education could be a major reason.

Attempts to compare relative returns across groups become particularly confusing and illuminate very little. It is worth noting that in regressions by Mwabu and Schultz (2000) and Hosking (2001) which are run separately by race, Africans are generally found to have higher returns to all levels of education. This does not imply that Africans earn more by education level, but rather that their increase in earnings per additional year of education is greater. The reason for this would be that when Africans invest in additional education, they forego only a low level of earnings, thus leading to returns that are higher as they are off a particularly low investment base. Higher returns to education for Africans would imply decreasing between race inequality. No consistent pattern emerges by gender.

It is worth cautioning on the interpretation and comparisons of such results. In terms of returns to education, when separate regressions are run by race, the Mincerian estimate refers to, for example, the return to one additional year of schooling, if one is African, relative only to other individuals who belong to this race group. When separate regressions are run and then compared, total non-correlation of both parameters and the error terms is assumed. In earnings function regressions where the R^2 is typically in the region of 40%, most variation in earnings remains unexplained. If there are similar processes driving this unexplained variation, then this covariance between the error terms should be accounted for. In the case of separate earnings functions, it is likely that between group comparisons yield inconsistent results because they are not based on a correct "between group mean". Alternative techniques that might better capture probable covariance across error structures (as well as allowing for flexible parameter estimates for each group) include estimating a single equation and including dummy variables for race interacted with the education term or using seemingly unrelated regression (see Griffiths, W.E., Carter Hill, R. and G.G Judge (1993), chapter 17).

In this section we have reviewed evidence on returns to education for South Africa and the rest of the world. We have seen that even though original studies found returns to be diminishing, there

[10] Factors such as emigration also contribute to the skill's shortage.

is no *a priori* reason to expect that this is any longer the case. Indeed, the pattern in South Africa appears to be increasing, a result that holds true for other African countries we have considered.

5. Empirical Analysis

The literature review revealed that the pattern of returns to education in South Africa are most probably convex. In the remainder of the paper we use empirical methods in an attempt to verify these findings and gain a more detailed understanding of the pattern of returns to schooling in the country. We run a number of earnings functions on 4 nationally representative datasets to ascertain whether the apparent convexities are a robust feature of the data. We also employ semi-parametric techniques to explore the relationship between education and earnings in more detail.

5.1 Data

The original surveys selected for the study were the October Household Surveys for 1995, 1997, 1998 and 1999[11], the Project for Living Standards or PSLSD for (1993), the KwaZulu-Natal Income Dynamics or KIDS Surveys (1993 and 1998) and the Labour Force Survey (September 2000). Owing to completeness of data reported and representativeness of the surveys, we report results for 4 of the surveys here, namely the PSLSD, 1993, OHS, 1995, OHS, 1997 and LFS, 2000.

In the simple Mincerian earnings function, the key variables of interest are education, earnings and age (or experience). In our analysis we define education as completed years of schooling. Earnings is measured as 'total monthly pay including overtime and bonuses and before any tax or other deductions'[12]. The employed are restricted to full-time wage employees. The unemployed are defined according to the broad definition, in which discouraged work-seekers are included. Other covariates are race, gender, union status and province. Appendix B details issues concerning the definitions of the employed, the unemployed (or censored observations) and earnings, as well as, the representativeness of the final datasets used in our study.
Below is a table of descriptive statistics based on the final sample used for each of the 4 datasets that are carried through this analysis.

[11] The OHS, 1996 was excluded primarily due the nature of the income data which were only reported in categories. OHS, 1996 is also least comparable with other years as its sample size is approximately half that of the other years and the enumerator areas were selected on a different basis. OHS, 1994 was also excluded as it was in a sense the "pilot OHS" and also had only 1000 enumerator areas as opposed to the 3000 for the surveys of 1995, 1997, 1998 and 1999.
[12] Although net earnings after tax would have been the preferred variable, no data was collected on income tax paid in the OHS, 1997 or the LFS, 2000 (as well as in the OHS, 1998 and OHS, 1999).

Table 5: Descriptive Statistics reflecting the final sample analysed for PSLSD, 1993, OHS, 1995, OHS, 1997 and LFS, 2000

	PSLSD, 1993	OHS, 1995	OHS, 1997	LFS, 2000	Actual EAP[a]
Economically active population (EAP) under consideration (sample size)	8495	31777	30956	29949	Not applicable
Race breakdown					
African	76.14	70.22	79.61	79.76	**72**
Coloured	10.18	14.70	14.32	12.48	**11**
Indian	3.64	3.71	1.84	2.16	**3**
White	10.04	11.37	4.23	5.60	**14**
Gender					
Male	52.48	56.06	49.44	49.51	**54**
Female	47.52	43.94	50.56	50.49	**46**
Location					
Rural	77.93	42.74	44.07	39.07	**36**
Urban	22.07	57.26	55.93	60.93	**64**
% of Employed unions members					
Not Union	71.2	67.59	62.14	60.97	
Union member	28.8	32.41	37.86	39.03	
Average Age	32.88	34.90	34.07	34.18	
	(11.10)	(11.09)	(10.63)	(11.16)	
mean age of censored observations	30.12	31.25	31.34	30.56	
	(10.64)	(10.31)	(9.96)	(10.33)	
mean age of employed	35.70	37.41	37.19	38.52	
	(10.85)	(10.90)	(10.51)	(10.54)	
Average Educational attainment	7.58	8.23	7.82	8.50	
	(3.99)	(3.99)	(3.97)	(3.81)	
mean education of censored observations	7.07	7.66	7.70	8.42	
	(3.72)	(3.75)	(3.78)	(3.54)	
mean education of employed	8.10	8.62	7.94	8.60	
	(4.19)	(4.09)	(4.17)	(4.10)	
EAP					
Censored	50.58	40.68	53.26	54.52	
Employed	49.42	59.32	46.74	45.48	
Average income for employed					
Mean income of employed	1723.46	1934.35	1632.10	2443.86	
	(3464.60)	(2231.36)	(1734.08)	(3567.08)	
Mean log income of employed	6.87	7.08	6.93	7.28	
	(1.12)	(1.04)	(1.03)	(1.04)	
Median	1000.00	1268.00	1200.00	1500.00	

Notes:
[a] The 'Actual EAP' statistics are the average proportions by race, gender and location of the economically active population calculated from the weighted sample surveys (PSLSD 1993, OHS, 1995, OHS, 1997 and LFS, 2000). They therefore indicate what the breakdown by race, gender and location should be given no missing information and totally representative sampling.
Standard errors are in parentheses.

It is important to note that all the surveys are nationally representative and the samples used consist of a large number of data points. The economically active population ranges from almost 8500 in PSLSD, 1993 to over 31 500 for OHS, 1995.

Given no missing income information, the proportion of the economically active population by race, gender and location should be similar to the values indicated in the final column. It is evident that OHS, 1995 most closely approximates the actual breakdowns for both race and gender. The OHS, 1997 and LFS, 2000 clearly under-represent Whites and over represent Africans. This is predominantly due to these surveys undersampling Whites and then weighting up the aggregates. OHS, 1997 and LFS, 2000 also reveal a higher proportion of economically active females. This could reflect a trend of increases in female labour force participation over the timespan under consideration (see Posel and Casale, 2001). The LFS, 2000 is the most representative by location[13]. There could also be a time effect picked up here reflecting high levels of annual rural to urban migration. Such apparent trends, however, may be a function of the sample design.

The percentage censoring ranges from 41 to 55 percent with OHS, 1995 once again revealing the most realistic level. The average age of the employed rises from a minimum of 36 years for PSLSD, 1993 to a maximum of 39 years in LFS, 2000. The average age of the censored observations is a fairly constant around 31. Increasing youth unemployment and low labour absorption rates would fit well with such a picture of a widening in the age gap between the employed and unemployed. So too would the apparent increase in mean educational attainment of the censored observations for which the statistics indicate has increased by 1 ½ years.

Average income for the employed under consideration increases from R1732 in 1993 to R2444 in 2000. Income data for 1993, 1995 and 2000 suggests that nominal earnings have increased by 6 percent per annum over the period. With average inflation of approximately 7% per year from 1993 to 2000, these seem like reasonable estimates indicating that their has been slight negative growth in real earnings over the period. The average income data for OHS, 1997 is lower due to the undersampling of Whites and Asians and those with higher education.

[13] The breakdown by location for the PSLSD, 1993 is quite different from the other surveys. This is largely because the PSLSD purposefully over-sampled in the rural areas.

The OHS, 1995 therefore seems the most representative survey and this should be kept in mind when analysing results. The LFS, 2000 is the only survey whose exclusive purpose was to explore the labour market and therefore data for labour market variables may be captured better than for the other surveys. The time dimension may play a role when comparing PSLSD, 1993 with later surveys as the labour dynamics in the last 7 years may mean that there have been changes in the covariates used in the analysis. Finally, OHS, 1997 seems to be the least representative of the surveys. This should be taken into account when analysing differences in outcomes between this and other datasets.

5.2 Estimation

In this section we develop the empirical model estimated and hypothesize on expected results. The aim of the empirical analysis was to estimate the **total** private returns to education thereby including the role of education in terms of influencing the probability of finding employment, as well as its effect on earnings. Censored observations are therefore included in all our models.

To begin with, the return is estimated using a parametric regression framework. A sample selection problem presents itself in that earnings are captured only for those who are employed. This situation lends itself to using a censored regression approach. We choose the tobit as our sample selection model in preference to the Heckman as identification of the Heckman with comparable variables across surveys would be an extremely arduous task. We also present OLS estimates for comparative purposes[14]. We ran ten different earnings functions for each of the four datasets using both the tobit and ordinary least squares estimators[15].

The dependent variable in all equations is the log of gross monthly earnings.[16] In the first equation education is brought in as a discrete term as "years of education" with "age" and "age squared" as the other explanatory variables – this is the Mincerian earnings function using age

[14] Although running OLS regressions on a censored sample should produce biased estimates (see Appendix C), the coefficients are more easily interpreted than for maximum likelihood models which are inherently non-linear. If the marginal effects for the different types of models do not differ widely, OLS estimates can aid in simplifying interpretations.

[15] The reason for running a number of different earnings functions was to check the consistency of the coefficients as further control variables were added.

[16] The choice of using monthly earnings as opposed to hourly is largely a function of the question dealing with "hours usually worked per week" that changes across surveys. As our sample of interest is limited to full time formal sector workers, this is not seen as particularly problematic. Furthermore, if one assumes

instead of potential experience which is estimated according to Mincer (1974) as *"Age-Education-6"*. [17]. "Age" is used in preference to "experience" as the Mincerian proxy is unreliable for a country such as South Africa with a labour market characterized by high levels and long periods of unemployment and a schooling system with high repetition rates. In other equations we add higher order education terms to allow for any non-linearities in the returns to education. We also include "age-education" interaction terms to control for the differing average educational attainment of various age groups. These terms are especially important in that they circumvent a particular problem with interpreting the education coefficient as a return. The traditional Mincerian estimate overstates the opportunity cost of early years of education as young school going children would generally not be earning if they were not at school (do not forego the equivalent years of earnings), or would be earning substantially less than adults (Psacharopoulos, 1994:1326). The rate of return to low levels of education is therefore generally understated. Interaction effects control for this form of bias.

Single earnings functions are estimated with covariates such as race, gender, location or union status included as dummy variables. The single regression approach is taken so as not to assume total independence of the unobserved factors driving the returns to education for the full-time wage earners across subgroups. Between race comparisons are made later in the semi-parametric analysis in which race-education interaction effects are included.

Finally, it is acknowledged that there are many different kinds of biases working in different directions that may impact on our results. Indeed, Hertz (2001) has found problems of measurement error and omitted variable bias to be large for South Africa. The aim of this empirical work is to estimate a set of functions that are as consistent as possible, and subject to as little distortion as possible through additional econometric manipulation. Inclusion of any sophisticated variables across surveys in which the questionnaire design differs will introduce further noise into our results. Moreover, the lack of data on factors such as school quality and ability (for example test scores) across surveys limits the potential to control for these types of biases. It is chosen therefore only to control for sample selection bias. Given the high unemployment rate in the country, we believe it is important to include the effect of education on

---

that the pattern of hours worked over the lifecycle is fixed exogenously, it makes little difference what measure of earnings is used, be it hourly, weekly, monthly or annual (Willis, 1986).

[17] Even though 'years of education" is discrete rather than continuous, as it is roughly monotonic it is treated here as if it is continuous.

the probability of finding employment. The tobit is used to control for sample-selection effects, the formal model of which is as follows;

$$y_i = x_i'\beta + \mu_i \qquad \text{if } y_i > 0$$
$$y_i = 0 \qquad \text{otherwise}$$

where y_i is measured in our model as the log of monthly earnings, x_i is the matrix of relevant explanatory variables, β is the associated coefficients matrix and μ_i the stochastic error term, assumed identically and independently distributed.

For $y_i = 0$

We find $pr\,(y_i = 0) = pr(u_i < - x_i'\beta) = (1-\Phi_i)$

Which uses the result of cumulative distribution function of a standard normal random variable with mean zero and variance σ^2. (see equation (5) Appendix C)

For $y_i > 0$, we use the result for the standard normal density function evaluated at $x_i'\beta/\sigma$ or ϕ_i (see equation (8) Appendix C) and find

$$pr(y_i > 0) \cdot \phi(y_i \mid y_i > 0) = \Phi_i \frac{1}{\sigma} \frac{\phi[(y_i - x_i'\beta)/\sigma]}{\Phi_i} = \phi[(y_i - x_i'\beta)/\sigma]$$

Using maximum likelihood to solve we get the likelihood function

$$l = \prod_{y_i, y_i = 0} [1 - \Phi_i] \prod_{y_i | y_i > 0} \phi[(y_i - x_i'\beta)/\sigma]$$

where the first term is equivalent to the probit reflecting the probability of an observation being censored and the second term to OLS estimated on only uncensored observations.

The log likelihood function (shown below) is then maximised by taking the derivative with respect to β and σ.

$$L = \sum_{y_i : y_i = 0} \log(1 - \Phi_i) + \sum_{y_i | y_i > 0} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_1 \frac{1}{2\sigma^2}(y_i - x_i'\beta)^2$$

The values of β and σ that maximise the log likelihood do not have closed form solutions. They must therefore be found through numerical methods. This also implies that the marginal effects of particular explanatory variables have to be evaluated at relevant values of x_i, the most common being evaluation at the means.

Most of our discussion of the results will be based on the final specification in which all control variables referred to above are included. It is expected that the coefficient on female will be negative relative to male and that union members will earn a premium relative to non-union members. Africans are expected to be the most disadvantaged relative to whites, with Coloureds and Asians also exhibiting lower earnings. All provinces other than the Western Cape are likely to have negative coefficients relative to Gauteng. The concave age earnings profile should reveal a positive coefficient on the "age" term and a negative on the "age squared" term. We expect schooling to have a positive impact on log earnings and study the coefficients in detail to reveal the pattern of returns by level of educational attainment.

5.3 Results

Regressions were run using both the tobit and OLS including the zero earners and it was found that the OLS and evaluation of the tobit at the means of the explanatory variables yielded (surprisingly) similar results. Referring to Table 6, Table 7 and Appendix D the regression results for both the OLS and tobit reveal that the coefficients tend to have the expected signs. The concave age-earnings relationship is evident with the "age" coefficient taking a positive sign and "age^2" a negative. Union membership contributes positively to earnings as does being white relative to the other races. The province dummies have the expected signs for 1995 and 2000 with coefficients negative with respect to "Gauteng" except for the "Western Cape" which is insignificant in 1995 and positive in 2000. An unexpected result emerges with positive and significant coefficients on the province dummies for "Northwest" and "Mpumulanga" in PSLSD, 1993 and for "Northern Cape", "Free State" and "Mpumulanga" for OHS, 1997. Closer inspection of the data reveals that for both datasets the percentage censoring in "Gauteng" is considerably higher than that for the above-mentioned provinces[18]. It is therefore the inclusion of the censored observations and the probability of finding employment that is driving this result.[19]

We now concentrate our discussion on the role of education.

Looking closely across the datasets at Specification 1, which is the basic Mincerian earnings function, it can be seen that the coefficient on the "years of education" variable ranges from 17 to 26.4 % for OLS and from 15 to 26.6% for the tobit at the mean of the explanatory variables. This result will be higher than findings where only the employed are included as the impact of education on the probability of finding employment is incorporated in the result. (Hertz (2001) for example finds an average coefficient of 0.114 for African males in the PSLSD, 1993 when he includes only the employed. Hosking (2001) finds a coefficient of 0.141 for Africans using the 1996 Census and a coefficient of 0.189 when running the Mincerian earnings function for all races.)[20] The coefficient of determination (R^2) ranges from 12% to 19% indicating the effect of education and age in explaining the variation in log earnings is large and significant. This implies that differences in education are an important determinant of earnings inequality. Including higher order education terms into the model leads to the R^2 being on average 2.8 percentage points higher indicating better explanatory power with the more flexible functional form.

[18] The "effective rate of unemployment" for Gauteng is 52% as opposed to 40% for Northwest and 37% for Mpumalanga for SALDRU93. In the OHS97 the percentage censoring is 54% for Gauteng as opposed to 37%, 45% and 49% for Northern Cape, Free state and Mpumalanga respectively. It should be noted that the percentage cnesoring rates are the rates on which our data is based and not necessarily a reflection of the actual relative unemployment rates per province.

[19] Initially a dummy variable controlling for location (urban=1, rural=0) was included but in specification 9 the effect of being in an urban area was negative for all datasets and for both OLS and the tobit. The unemployment rate of urban areas was however lower than for rural areas and mean income of the entire labour force also a great deal higher in urban areas. Through further testing it appears that including the unemployed in the regression introduces a degree of covariation with regards to the right hand side variables. "Urban" covaries with education, race and union status. As inclusion of "urban" biases results, we run our regressions excluding it.

[20] The inclusion of only full time wage employees as the 'employed' should also serve to drive estimates of returns to higher levels than if part time and temporary workers were additionally part of the sample considered.

Table 6: The Mincerian Earnings Equation: Specification 1 – Marginal effects for OLS and Tobit estimators

	OLS93	Tobit 93	OLS95	Tobit 95	OLS97	Tobit 97	OLS00	Tobit 00
Constant	-5.713 *	-9.750 *	-5.751 *	-9.588 *	-4.923 *	-9.190 *	-7.546 *	-12.865 *
	(0.319)	(0.378)	(0.177)	(0.216)	(0.201)	(0.235)	(0.179)	(0.215)
Age	0.339 *	0.407 *	0.336 *	0.409 *	0.292 *	0.366 *	0.399 *	0.517 *
	(0.018)	(0.021)	(0.009)	(0.011)	(0.011)	(0.013)	(0.010)	(0.011)
Age Squared	-0.003 *	-0.004 *	-0.003 *	-0.004 *	-0.002 *	-0.003 *	-0.003 *	-0.005 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	0.245 *	0.232 *	0.263 *	0.265 *	0.171 *	0.152 *	0.202 *	0.182 *
	0.009	(0.01)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)
n	8487	8487	31777	31777	30956	30956	29949	29949
R^2	0.162		0.180		0.121		0.192	
Log likelihood		-16352.6		-68045.2		-58284.3		-54712.2

Table 7: Specification 10 – Marginal effects for OLS and Tobit Estimators (Higher order education terms and controlling for all other covariates)

	OLS93	Tobit 93	OLS95	Tobit 95	OLS97	Tobit 97	OLS00	Tobit 00
Constant	1.091 *	-2.9270 *	0.560 **	-3.037 *	1.019 *	-3.060 *	-0.021	-5.187 *
	(0.412)	(0.510)	(0.244)	(0.321)	(0.261)	(0.317)	(0.288)	(0.359)
Age	0.208 *	0.2750 *	0.247 *	0.328 *	0.174 *	0.238 *	0.238 *	0.344 *
	(0.018)	(0.022)	(0.010)	(0.012)	(0.011)	(0.013)	(0.010)	(0.013)
Age Squared	-0.002 *	-0.0030 *	-0.002 *	-0.003 *	-0.002 *	-0.002 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000))	(0.000)	(0.000)
Female	-0.545 *	-0.5770 *	-1.362 *	-1.716 *	-0.837 *	-0.918 *	-0.682 *	-0.758 *
	(0.058)	(0.070)	(0.031)	(0.039)	(0.031)	(0.037)	(0.031)	(0.037)
African	-3.165 *	-2.8380 *	-2.498 *	-2.571 *	-2.293 *	-2.019 *	-2.597 *	-2.243 *
	(0.113)	(0.127)	(0.055)	(0.066)	(0.082)	(0.089)	(0.073)	(0.078)
Coloured	-2.117 *	-1.6420 *	-1.257 *	-1.079 *	-1.186 *	-0.785 *	-1.504 *	-1.048 *
	(0.154)	(0.173)	(0.067)	(0.081)	(0.095)	(0.103)	(0.088)	(0.096)
Asian	-0.894 *	-0.5390 **	-0.535 *	-0.434 *	-0.206	0.188	-0.622 *	-0.365 *
	(0.189)	(0.211)	(0.093)	(0.111)	(0.138)	(0.151)	(0.128)	(0.138)
Education in years	0.220	-0.0650	0.841 *	0.929 *	0.547 *	0.595 *	0.485 *	0.442 *
	(0.167)	(0.207)	(0.096)	(0.125)	(0.101)	(0.121)	(0.106)	(0.129)
(Education in years)^2	-0.089 *	-0.0500	-0.191 *	-0.215 *	-0.130 *	-0.150 *	-0.106 *	-0.107 *
	(0.031)	(0.038)	(0.017)	(0.021)	(0.018)	(0.021)	(0.017)	(0.02)
(Education in years)^3	0.006 *	0.0050 **	0.010 *	0.012 *	0.007 *	0.008 *	0.005 *	0.006 *
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age* Education	-0.005	0.0010	-0.021 *	-0.024	-0.015 *	-0.017 *	-0.011 *	-0.011 *
	(0.004)	(0.005)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Age* Education^2	0.002 **	0.0010	0.004 *	0.005 *	0.003 *	0.004 *	0.002 *	0.002 *
	(0.001)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Age* Education^3	0.000 **	0.0000 ***	0.000 *	0.000 *	0.000 *	0.000 *	0.000 *	0.000 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Union	3.232 *	2.9060 *	3.043 *	3.140 *	4.033 *	3.436 *	4.324 *	3.545 *
	(0.085)	(0.094)	(0.040)	(0.048)	(0.043)	(0.047)	(0.044)	(0.048)
Western Cape	1.132 *	1.1310 *	-0.041	0.019	0.920 *	0.942 *	0.408 *	0.393 *
	(0.136)	(0.157)	(0.068)	(0.083)	(0.072)	(0.083)	(0.073)	(0.083)
Eastern Cape	-0.349 *	-0.5590 *	-1.068 *	-1.281 *	-0.800 *	-1.019 *	-0.924 *	-1.115 *
	(0.112)	(0.142)	(0.057)	(0.072)	(0.060)	(0.076)	(0.060)	(0.073)
Northern Cape	0.007	0.0480	-0.594 *	-0.598 *	0.416 *	0.518 *	-0.168 ***	-0.083 *
	(0.260)	(0.305)	(0.089)	(0.112)	(0.088)	(0.099)	(0.087)	(0.099)
Free State	-1.024 *	-1.8790 *	-0.238 *	-0.086	0.437	0.584 *	-0.340 *	-0.292 *
	(0.170)	(0.252)	(0.063)	(0.078)	(0.062)	(0.073)	(0.065)	(0.075)
KwaZulu-Natal	0.009	-0.0570	-0.358 *	-0.396	-0.493	-0.672 *	-0.402 *	-0.453 *
	(0.102)	(0.126)	(0.055)	(0.068)	(0.054)	(0.066)	(0.054)	(0.064)
Northwest Province	1.038 *	1.2330 *	-0.269 *	-0.258 *	0.032	0.066	-0.388 *	-0.424 *
	(0.114)	(0.135)	(0.067)	(0.083)	(0.060)	(0.072)	(0.059)	(0.07)
Mpumulanga	1.181 *	1.3760 *	-0.368 *	-0.406 *	0.332 *	0.423 *	-0.520 *	-0.597 *
	(0.115)	(0.135)	(0.064)	(0.08)	(0.063)	(0.074)	(0.066)	(0.079)
Northern Province	0.101	0.1160	-0.869 *	-1.075 *	-0.096	-0.063	-0.785 *	-0.947 *
	(0.118)	(0.144)	(0.070)	(0.09)	(0.064)	(0.077)	(0.062)	(0.076)
n	8487	8487	31777	31777	30956	30956	29949	29949
R^2	0.449		0.441		0.428		0.474	
Log likelihood		-14985		-62975		-53247		-50255

Notes:

Table 6 and Table 7 report estimates for both ordinary least squares and the tobit (derived above) where the dependent variable is the log of monthly earnings from full time formal sector workers. Given the non-linearity of the tobit, the estimates reported are the marginal effects of the model evaluated at the mean of the relevant explanatory variables according to the formula $\frac{\partial E(y)}{\partial x_j} = \Phi(z)\beta_j$

* implies statistical significance at the 1% level, **implies significance at the 5% level, and *** implies significance at the 10% level. Standard errors are in parentheses.

On adding further controls to the basic Mincerian equation it can be seen that the coefficient on education reduces quite substantially (refer to specifications 8 and 9, Appendix D). This indicates that the education term and other right hand side variables are associated (Griliches and Mason 1972). Through deeper analysis it is found that for the OHS, 1997 and the LFS, 2000, it is largely the inclusion of the union dummy that is driving this result whereas for the PSLSD, 1993 and the OHS, 1995 it is the race effect[21]. Omission of these variables then biases education estimates upwards. On inclusion of the controls for race and union, the returns to education at the mean are extremely stable for both the tobit and OLS, across the different functional forms and when other controls such as gender and province are added. Furthermore, the calculated total return of education at the means for specification 10 is very similar to the (average) coefficient on the single "years of education" term for specification 9 (including race, gender and union: see Appendix D). This establishes that our results are consistent across years and specifications once race and union effects are included.

Considering the higher order education terms reveals a particularly robust result across specifications and datasets: The signs of the coefficients are positive on the first order term, negative on the second order and again positive on the third order indicating an S-shaped relationship between education and earnings. The only exception to this pattern is the tobits using the PSLSD, 1993. It is not surprising that the earnings-education relationship seems to differ for PSLSD, 1993. This may be a function of survey design or in fact a time dimension. We are dealing with a quite different generation of school leavers in 1993 and 2000[22].

To calculate the marginal effect of education from the equations with higher order terms, it is necessary to take the derivative of log earnings with respect to education as follows.
Given the equation

$$Lnwage = \alpha + \beta Age + \chi Age^2 + \delta Education + \varepsilon Edu^2 + \gamma Edu^3 + \eta(Age \times Edu) + \kappa(Age \times Edu^2) + \lambda(Age \times Edu^3) + \ldots..$$

Take the derivative to get

[21] Further regressions were run with different combinations of variables to establish at which point the coefficients changed.
[22] Rerunning the regressions for PSLSD, 1993 with only a square and cubic term for education and the age-education interactions, yields highly significant coefficients on both these terms. Predicted returns from these regressions make little difference to those including the "years of education" term in the final result.

$$\frac{\partial Lnwage}{\partial Education} = \delta + 2\varepsilon Education + 3\gamma Education^2 + \eta Age + 2\kappa(Age \times Ed) + 3\lambda(Age \times Ed^2)$$

The marginal effects can then be calculated for a specific level of education at a given age by substituting into the equation above. The tables below shows the marginal returns for a 40 year old at 7, 10, 12 and 15 years of education and the average returns per year of primary, secondary and tertiary education. These results are taken from specification 10.[23]

Table 8: Marginal Effect of Education for a 40 year old

Years of Schooling	OLS93	Tobit93	OLS95	Tobit95	OLS97	Tobit97	OLS00	Tobit00
7	0.09	0.08	0.06	0.06	0.04	0.03	-0.01	-0.02
10	0.28	0.27	0.27	0.29	0.20	0.19	0.15	0.14
12	0.46	0.46	0.46	0.53	0.35	0.36	0.32	0.32
15	0.80	0.89	0.85	1.05	0.65	0.76	0.54	0.61

Table 9: Average Effect of an additional year of Education for a 40 year old

Years of Schooling	OLS93	Tobit93	OLS95	Tobit95	OLS97	Tobit97	OLS00	Tobit00
3 - 7	0.02	0.02	0.00	-0.01	-0.02	-0.03	-0.04	-0.06
8 - 12	0.29	0.28	0.28	0.30	0.21	0.20	0.16	0.15
13 - 15	0.68	0.73	0.71	0.86	0.54	0.61	0.54	0.61

Notes:
The average effect is calculated by averaging the marginal effect for each year of the relevant education span

The above result shows that across datasets the pattern of returns is increasing. We see the previously discussed very low returns to 7 years of education as found by most other research on South Africa. We then see the returns increasing to high levels at 12 years and extremely high levels at 15 years. Such a pattern of returns is clearly indicative of a convex relationship between education and earnings.[24]

[23] In order to calculate the correct coefficients for the tobit, the marginal effects were re-estimated for each given year of education with age specified at 40 and all other variables held at their means. This should give the marginal effect for the "average 40 year old" in the sample.

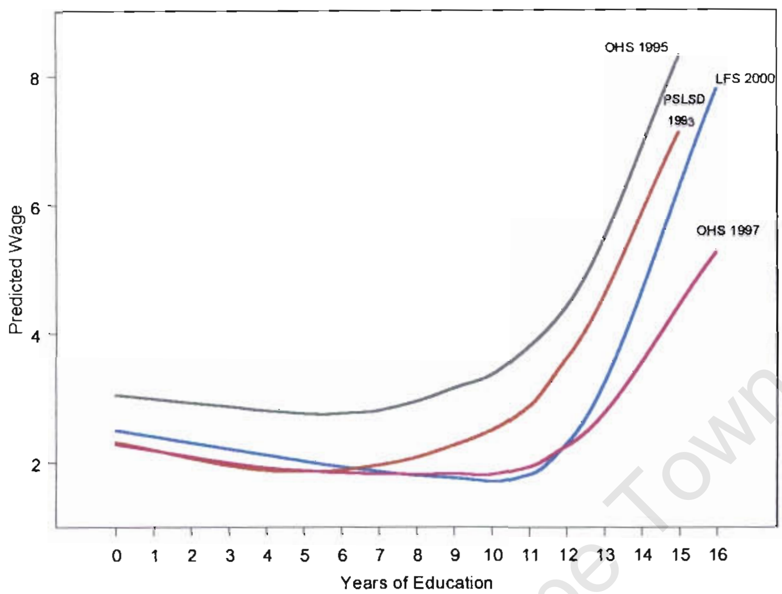
[24] The returns to an additional year of higher education do however seem particularly high. Hertz's (2001) previously mentioned finding of extreme upward bias in the Mincerian coefficients indicates that the actual return estimates should in all probability be much lower. However, it is predominantly the level of returns that change on correction, with the pattern remaining fairly stable. As the focus of our exercise is the pattern of returns, we do not discuss the levels of returns in any detail. It is important to notice, however, that the absolute returns vary substantially across datasets. This finding warns at drawing conclusions with reference to the size of returns when applying basic earnings function techniques to the datasets studied here.

Although the pattern of increasing returns found above is robust across datasets and appears plausible for a country like South Africa, parametric analysis still constrains the data to quite specific relationships. There may be turning points that the higher order education terms are not in fact capturing. To better visualize the relationship between log earnings and education, it is useful to employ semi-parametric techniques. These will potentially reveal turning points or other characteristics of the data that may be concealed when using purely parametric methods.

The non-parametric approach we use is the LOWESS method or Locally Weighted Scatter Plot Smoothing. LOWESS works by starting with a local polynomial least squares fit and then iteratively identifies weights and resmoothes a number of times (Hardle (1995:192-193)). In LOWESS a separate weighted regression is run for every datapoint. The estimation process weights the values closest to a particular point the highest and those furthest away the least. The local nature of the regressions means that LOWESS is particularly useful in the presence of outliers (Stata Reference Manual, 2001:170-171). The approach we use is semi-parametric in that the observations on which the LOWESS is performed are the predicted values of the natural log of earnings for each year of education obtained from the parametric regressions discussed above. The reason we use the predicted as opposed to the actual values, is that in this way the effects of the other covariates have been to some extent controlled for.

We view the results below. Figure 1 shows the LOWESS plot for "predicted log earnings" by "years of education", for all individuals in each dataset. The slope of the plots reflects how predicted earnings change with each level of education. That is, it represents the private return.

Figure 1: The relationship between education and predicted earnings for all datasets

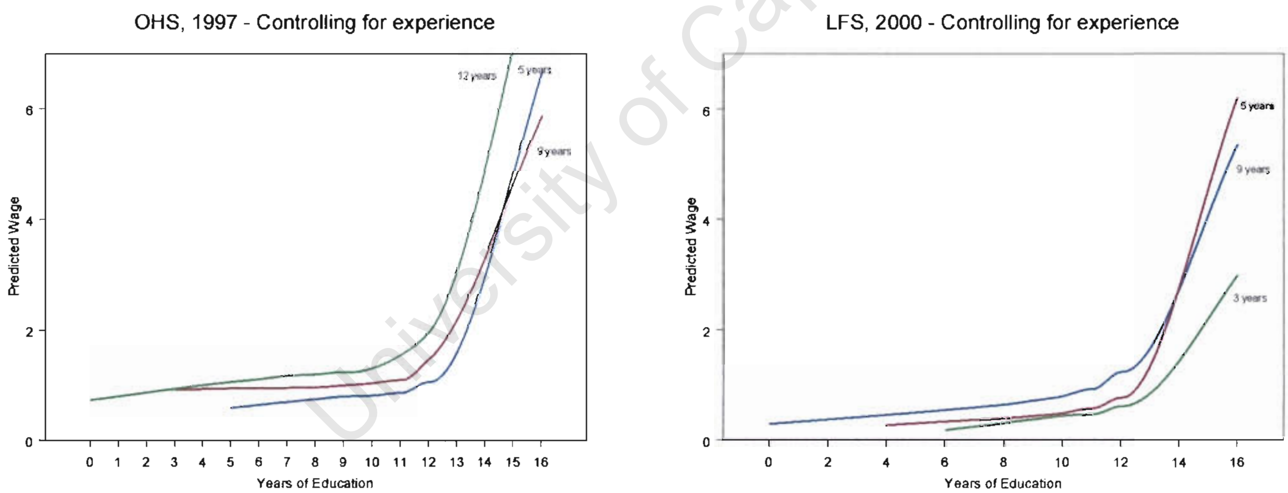


Notes:
The vertical axis is the log of predicted earnings, controlling for age, age^2, gender, race, years of education (including squared and cubic), age-education interactions, and regional fixed effects. The predicted values were generated using the Tobit estimator, to control for sample selection bias introduced through conditioning on whether an individual was employed or not, and therefore whether she reported earnings or not. The full sample including the zero earners is shown in Table 5 above. The pictures are generated through applying non-parametric Locally Weighted Scatter Plot Smoothing to the underlying fitted regression values against years of reported education, a method that is robust to outliers.

If we look firstly at the predicted wage for all respondents for each dataset we clearly see the convex relationship between education and earnings. The predicted wage actually declines, although only slightly, up until approximately 7 years for the PSLSD, 1993 and OHS, 1995 and until 10 years of education for OHS, 1997 and LFS, 2000. All years under consideration rise sharply from 11 or 12 years[25]. It is important to note how flat the slope is until the 11 year mark, indicating the very small contribution of additional education to the predicted wage for all of primary and secondary schooling up to matric. The flatness of the slope is equivalent to the low, close to zero, returns that are generally reported for primary education. We see from these diagrams that these low returns actually appear to extend all the way through to having completed 8 to 11 years of education, depending on the dataset.

Hertz (2001:54) finds the downward sloping graphs a robust feature of his regressions that are run for African males in the PSLSD, 1993 and have monthly earnings as the dependent variable. He offers two explanations for this result. First, he refers to potential importance of basic literacy (that is achieved at low levels of education) in helping to find employment. Second, he argues that beyond the basic literacy level of schooling there is likely to be a dominant labour market experience effect. This could occur if the majority of people with lower education are older. The apparent higher return to primary education may then be a function of age or labour market experience, rather than education level. He finds that on controlling explicitly for potential years of experience and thus capturing the trade off between experience and schooling, that the relationship between years of schooling and predicted earnings becomes positive. We perform the same test for our two datasets in which the downward-sloping effect is most severe. The graphs below show that for OHS, 1997 and LFS, 2000 on controlling for potential experience, the downward sloping portion becomes positive.

Figure 2: Predicted earnings and education controlling for potential experience



Notes:
The vertical axis is the log of predicted earnings, controlling for age, age^2, gender, race, years of education (including squared and cubic), age-education interactions, and regional fixed effects. The predicted values were generated using the Tobit estimator, to control for sample selection bias introduced through conditioning on whether an individual was employed or not, and therefore whether she reported earnings or not. The pictures are generated through applying non-parametric Locally Weighted Scatter Plot Smoothing to the underlying fitted regression values against years of reported education, a method that is robust to outliers. In controlling for potential experience, the LOWESS plot is generated for those individuals with the indicated years of potential experience only. Potential experience is calculated as "age"-"years of education"- 6 according to Mincer's (1974) method.

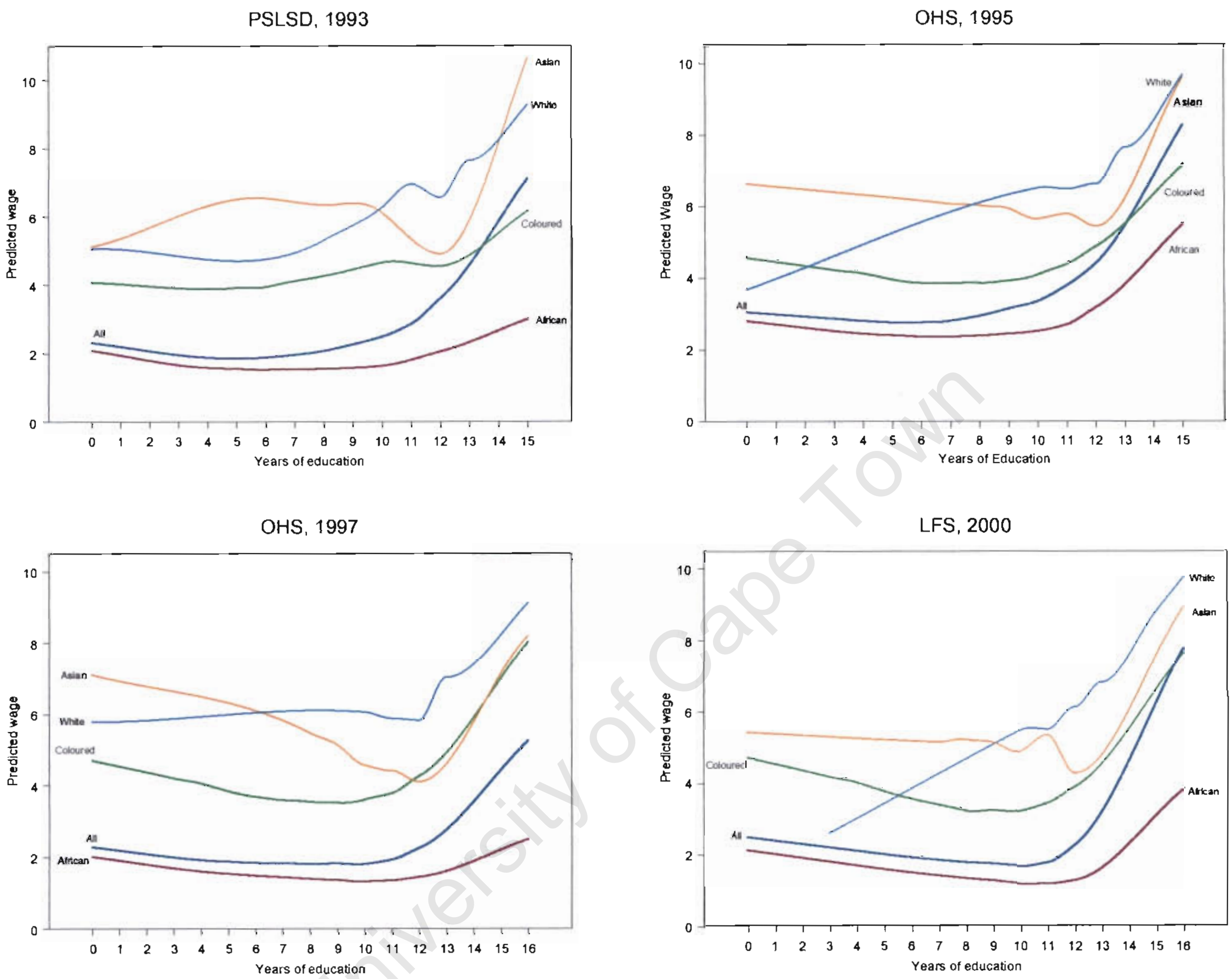
[25] It can be seen that predicted earnings vary quite considerably across datasets. This too cautions on drawing conclusions on the size of returns from such econometric analyses. The convex pattern, however, is found to be robust.

Hertz (2001) also examines the distribution of education misreporting and finds most errors occur when schooling is reported at three years or less. He therefore attributes the negative relation at low levels to data errors as opposed to labour market experience effects which he says are nonsensical as an explanation for young labour market participants that have not yet had time to forego accumulation of years of experience.

It is also important to notice that after the turning point, the graphs are in general rather straight and particularly steep. It appears that the returns are constant up until approximately 11 years, increase substantially around this point, and thereafter are fairly constant but at a much higher level. The LOWESS plots then reveal that returns do not appear to increase continuously as implied by the results of the parametric regression analysis. The pattern of returns may rather be described as "lumpy" or discontinuous. For example, the diagrams show that one could invest for 10 years, with practically zero return, or invest for more than 13 years and receive a high return. Such a pattern might more accurately be defined as a non-convex. This is precisely the type of situation modelled by Bardhan & Udry (1999) and Galor & Zeira (1993) discussed in Section 2. The implication of the non-convexity was the possible long run persistence of inequality.

We now consider race effects more closely. Figure 3 shows the LOWESS plot for "predicted log earnings" by "years of education" and by race group. Constructing separate graphs for each race group in this manner is equivalent to running a parametric regression that includes race-education interaction terms. Viewing the relationships in this way allows for analysis of both the within *and* between race effects. We can examine the between race effects because the underlying regressions on which the plots are based include all the races and use the between-group mean.

Figure 3: The relationsbip between education and predicted earnings including race effects



Notes:
The vertical axis is the log of predicted earnings, controlling for age, age^2, gender, race, years of education (including squared and cubic), age-education interactions, and regional fixed effects. The predicted values were generated using the Tobit estimator, to control for sample selection bias introduced through conditioning on whether an individual was employed or not, and therefore whether she reported earnings or not. The full sample including the zero earners is shown in Table 7 above. The pictures are generated through applying non-parametric Locally Weighted Scatter Plot Smoothing to the underlying fitted regression values against years of reported education, a method that is robust to outliers. The pictures indicate both with-in group and between group heterogeneity in the returns to education which is represented by both the curvature of the lines and the distance between them, respectively.

For both whites and Asians, there are so few observations for educational attainment of up to completed primary schooling that these should be ignored. One can interpret these graphs only from around 8 years[26]. For 1995, 1997 and 2000 we see that for whites the predicted wage increases with each additional year of education with a slight kink around the 12 year mark. The PSLSD, 1993 has the predicted wage increases starting at a later stage but also show the kink at matric. This kink may reflect short term frictional unemployment for school leavers searching for their first job. It is interesting that even from 8 years of education, the graphs for Africans and coloureds seem to exhibit a similar shape but those for Whites are quite different, at least until the turning points for the other races. The reason for this must be attributed to the low number of whites that are unemployed, thereby not pulling the return to education for this group to zero. For 1993, 1995 and 2000, the graphs for African and coloureds are practically parallel. This would infer that the processes affecting the returns to these two races are indeed very similar. The difference arises in that coloureds earn a significant premium. The slopes of the graphs for tertiary education appear in general to be steeper for Asians and whites, indicating higher returns to tertiary education for these two groups. Although the pattern fluctuates somewhat across years, there in fact appears to be convergence between the returns to education for these two races, at least for levels of tertiary education.

The relationships that are revealed in the above figures highlight the value of using the semi-parametric approach. In this way both the levels of predicted wages for the different races can be compared, as well as the returns that are based on the common group mean.

Our analysis of the plots serves to deepen our understanding of the relationship between education and earnings. We have corroborative evidence of the low returns to primary education indicated by the zero slope of the graph for years up to 7. We see this trend persist until around the 10 year mark. After the turning point it then appears that there is a much higher return to tertiary levels of schooling, but that this in itself is fairly constant. These findings differ somewhat from the parametric result that indicates returns increasing strongly from 8 or 9 years. The semi-parametrics have revealed an important turning point. The pattern of returns to different levels of education is found once again, to be increasing even if not as sharply for years post 13. We have also been able to consider the returns by race. The returns are found to be increasing within each race group. Between races it appears that it is whites and Asians who receive the greatest returns to tertiary education.

[26] The graph for whites for SALDRU93 is interesting in that there are a fair number of observations for those with zero education, and then very few until 8 years.

Once again the convexity of returns imply the possibility of poverty traps and persistent within group inequality. This result has been considered by both Lam (1999) and Hertz (2001) who assert that within-group increasing returns can be expected to lead to increased within-race inequality. Hertz comments further that between-race inequality is likely to decrease. In the sense that the distributions within race may become more similar even though more unequal, this view might hold. Indeed, Moll (2000) finds a decrease in between-race earnings inequality from 1980 to 1993 but an increase in within-race inequality. The change in the pattern of inequality is attributed to an increase in occupational mobility and the average educational attainment of Africans as well as a decline in discrimination. The link between the change in average educational attainment and the pattern of returns to education should, however, not be ignored.

Moll (2000) also states that from his study one cannot predict how inequality will change in the future. It is here that the findings of our semi-parametric analysis become particularly informative. The patterns we find by race suggest that coloureds and Africans receive lower returns to higher education. This would in fact suggest widening inequality between the race groups. Once discrimination is largely removed, *ceterus paribus*, it would be unsurprising to see persistent or even increasing inequality between whites and Asians as a group and Africans and coloureds as another group. This result cautions at discounting the importance of race as an explanatory variable due to apparent declines in between race inequality. The quality of schools and tertiary institutions, as well as occupational streaming, is still quite specific to the various races and the effects working through the return to education could have an impact on inequality for a long time to come.

6. Conclusion

The main objective of this paper was to ascertain with confidence, the pattern of returns to education in South Africa. Once this is established, the implications for inequality can be more clearly understood. Through thorough analysis of past studies and empirical modelling using both parametric and semi-parametric techniques, we conclude that the relationship between education and earnings in South Africa is convex with returns to schooling increasing with the level of education attained. The discontinuity found in the leap of returns from primary to higher education, might even be described as non-convex. Furthermore from our empirical analysis, it appears on a more rudimentary note that the returns to higher education for Africans and coloureds are lower than the returns for Asians and whites.

It seems likely that this pattern of increasing returns is driven by three main factors that can be explained in terms of the supply of low skilled workers, the demand for highly skilled workers and the large quantity of surplus labour in the economy. The vast supply of workers with low skills has resulted in earnings for this group being particularly meagre. This serves to drive down the returns to primary education. On the other end of the education spectrum, the limited supply of highly skilled workers and therefore excess demand has resulted in a particularly high skills premium. This pushes up the returns to tertiary education. Lastly, including the unemployed in the analysis serves to lower returns at the levels of education for which the highest proportions of the unemployed are found. The major portion of surplus labour in the economy is found amongst those who have relatively low levels of education, thus contributing to the pattern of increasing returns.

The pattern of returns has implications for the incentives driving individuals' human capital investment decisions. Increasing returns suggest that if education was free for everyone, all people would strive to accumulate as much as possible, so as to take advantage of higher labour market rewards. If an individual is faced with borrowing constraints, however, he must make his investment decision in an imperfectly functioning market. As Hertz (2001) explains, if a person is financially constrained and knows he will not be able to afford higher education he might quite rationally decide to drop out of schooling at a very low level, as the returns received are minimal. The expected pattern of the distribution of education would then be bifurcated. Based on the assumption that education maps to income, the consequent distribution of wealth would also be expected to bifurcate and persist. This outcome suggests persistent inequality in which those with initial wealth can invest in human capital and reap high returns whereas those who are financially constrained and unable to borrow cannot invest, do not accumulate human capital and become caught in a poverty trap. Such a situation relates closely to the models of multiple equlibria and low intergenerational mobility constructed by Galor and Zeira (1993) and Bardhan and Udry (1999) described in Section 1. The South African story is reinforced along racial lines due to past, rather than current discrimination.

In South Africa the minimum legal requirement for school attendance is that a child must go to school from the first day of the school year in which he turns 7 to the last day of the school year in which he turns 15 or completes grade 9, whichever comes first (Department of Education (1996)). Our finding of increasing returns would then indicate that one might expect a large number of teenagers leaving school after acquiring 8 or 9 years of education or upon reaching 15. This in fact seems to be the case as across the surveys approximately 50% of labour market participants aged 30 or below have an educational attainment of 9 years or less indicating a fairly high drop out rate at these levels. With secondary education highly subsidised by the state and more freely available to all races now than in the apartheid era, we would also expect to see a mass of people attain 12 years. Approximately 30% of those below age 31 have secondary education (20% for PSLSD, 1993).

The implications of the above create a fairly discouraging picture in terms of labour mobility and the persistence of inequality. The problem seems to be rooted in the exceptionally high unemployment rates for all levels of education below tertiary, as well as the prohibitive cost of tertiary education facing individuals with borrowing constraints. It appears that large interventions will be necessary to lift people out of poverty. Without good levels of economic growth and employment creation as well as greater access to funding for tertiary education either through borrowing or other means, it appears that the multiple equilibrium outcome and high levels of inequality may well persist for a long time to come.

7. References

Appleton, S., Hoddinott, J. and J. Mackinnon (1996) "Education and Health in Sub-Saharan Africa", *Journal of International Development*, 8:3, pp. 307-339.

Ashenfelter, Orley and Krueger, Alan. (1994) "Estimates of the Economic Return to Schooling from a New Sample of Twins", *The American Economic Review*, 84:5, pp.1157-1173.

Bardhan, P. and C. Udry (1999) *Development Microeconomics*, Oxford University Press, New York.

Barham, V., Broadway, R., Marchand, M. and P. Pestieu (1994) "Education and the Poverty Trap", *European Economic Review*, 39, pp1257-1275.

Becker (1964) Human Capital: *A Theoretical and Empirical Analysis, with Special Reference to Education*, New York: National Bureau of Economic Research.

Behrman, J. and N. Birdsall (1983) "The Quality of Schooling: Quantity Alone is Misleading", *American Economic Review*, 73:5, pp. 928-55.

Ben Porath, Y. (1967) "The Production of Human Capital and the Life Cycle of Earnings", Journal of Political Economy, 75:4 Part 1, pp352-365.

Bhorat, H. (2000) "Wage Premia and Wage Differentials in the South African Labour Market", Development Policy Research Institute, University of Cape Town: Cape Town.

Bhorat, H. (2001a) "Labour Market Challenges in the Post-Apartheid South Africa: A Country Profile" ILO commissioned study. Processed.

Bhorat, H. and M. Leibbrandt (2001) "Correlates of Vulnerability in the South African Labour Market" in Bhorat, H., Leibbrandt, M., Maziya, M. Van der Berg, S. and I. Woolard. *Fighting Poverty: Labour Markets and Inequality in South Africa.* UCT Press: Cape Town

Breen, R. (1996). *Regression Models: Censored, Sample-Selected, or Truncated Data,* Thousand Oaks, California: Sage Publications

Card, David (1999) "The Causal Effect of Ed on Earnings", in O. Ashenfelter and D. Card, eds., *Handbook of Labour Economics*, Vol 3A Amsterdam: Elsevier Science Publishers.

Card, David and Alan B. Krueger (1992) "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy*, 100:1 pp.1-40.

Carnoy, M. (1995) "Rates of Return to education." pp364-369, in M. Carnoy, ed., International Encyclopaedia of the Economics of Education, Second Edition, Oxford: Pergamon.

Casale, D and Posel, D (2001) "Why are more Women looking for Work? A Study of Female Labour Supply in South Africa, 1995-1999", Draft Paper presented at the DPRU/FES Conference on Labour Markets and Poverty in South Africa, Muldersdrift, 15-16 November.

Case, Anne and Yogo, Motohiro (1999) "Does School Quality Matter? Returns to Education and the Characteristics of Schools in South Africa", *National Bureau of Economic Research Working Paper* No 7399

Colclough, Christopher (1982) "The Impact of Primary Schooling on Development: A Review", *World Development*, 10:3 pp.167-185.

Deardon, Lorraine (1999) "Qualifications and Earnings in Britain: How reliable are conventional OLS estimates of the returns to education?" *The Institute for Fiscal Studies Working Paper Series* No. W99/7

Department of Education (1996). *South African Schools Act*, Section 3. Pretoria

Dinkelman, Taryn and Farah Pirouz (2002) "Individual, Household and Regional Determinants of Labour Force Attachment in South Africa: Evidence from the 1997 October Household Survey", *South African Journal of Economics*, 70(5) pp. 865-891.

Edwards, Lawrence (2002). "Trade, Technology and Employment in South Africa", *Trade and Industry Monitor* 23 pp.11-15.

Erichsen, G and J Wakeford (2001). "Racial Discrimination in South Africa Before and After the First Democratic Election", *DPRU Working Papers*, 01:49.

Galor and Zeira (1993) "Income Distribution and Macroeconomics", *The Review of Economic Studies*, 60:1 pp35-52.

Griffiths, W.E., Carter Hill, R. and G.G Judge (1993) *Learning and Practising Econometrics*, John Wiley & Sons, Inc. United States of America.

Griliches, Zvi and Mason, William M (1972) "Education, Income and Ability", *Journal of Political Economy*, 80:3 Part 2. pp. S74-S103.

Griliches, Zvi (Jan 1977) "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica*, 45:1 pp.1-22.

Halvorsen, R and R. Palmquist (1980) "The Interpretation of Dummy Variables in Semilogarithmic Equations", *American Economic Review*, 70:3 pp 474-475.

Hanoch, G. (1967) "An Economic Analysis of Earnings and Schooling", *Journal of Human Resources*, 2, pp. 310-329.

Hansen, W.L. (1963) "Total and Private Returns to Investment in Schooling", *The Journal of Political Economy*, 71:2 pp.128-140.

Hardle, W. (1995) *Applied Nonparametric Regression*, Econometric Society Monographs No.19. Cambridge University Press: Cambridge.

Hertz, T N (2001). "Education, Inequality and Economic Mobility in South Africa", *Department of Economics*. University of Massachusetts Amherst.

Hofmeyer, Julian (2001) "The Importance of Segmentation in the South African Labour Market", Paper presented at DPRU/ FES Conference on Labour Markets and Poverty in South Africa: Misty Hills, Johannesburg.

Hosking, S (2001) "Rates of Return to Education in South Africa, 1960-1996", Paper presented at DPRU/ FES Conference on Labour Markets and Poverty in South Africa: Misty Hills, Johannesburg.

Kane, T., Rouse, C.E. and D. Staiger (1999) "Estimating Returns to Schooling when Schooling is Misreported", National Bureau of Economic Research Working Paper No 7235.

Keswell, M. (2001) "Intragenerational Mobility: A Study of Chance and Change in Post-Apartheid South Africa", School of Economics, University of Cape Town.

Keswell, M and L. Poswell (2002) "How Important is Education for Getting Ahead in South Africa?", Centre for Social Science Research Working Paper No. 22, University of Cape Town.

Kingdon, G. and J. Knight (1999) "Unemployment and Wages in South Africa: A Spacial Approach", Centre for the Study of African Economies, Institute of Economics and Statistics, University of Oxford.

Kingdon, G. and J. Knight (2000) "Are Searching and Non-searching Unemployment Distinct States when Unemployment is High? The Case of South Africa." Centre for the Study of African Economies, Institute of Economics and Statistics, University of Oxford.

Lam, D. (1999) "Generating Extreme Inequality: Schooling, Earnings and Intergenerational Transmission of Human Capital in South Africa." Population Studies Center, ISR, University of Michigan: Ann Arbor.

Maddala, GS (1999) *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press: Cambrigde.

Michaud, P C and D Vencatachellum (2001) "The Union Wage Premium for Blacks in South Africa", Paper presented at DPRU/ FES Conference on Labour Markets and Poverty in South Africa: Misty Hills, Johannesburg.

Moll, P. (1996) "The Collapse of Primary Schooling Returns in South Africa, 1960-1990." *Oxford Bulletin of Economics and Statistics*, 58, pp. 185-209.

Moll, P. (2000) "Discrimination is declining in South Africa but Inequality is not", *Studies in Economics and Econometrics*, 24:3, pp 91-108.

Mwabu, G. and T. P. Schultz (1996) "Education Returns Across Quantiles of the Wage Function: Alternative Explanations for Returns to Education by Race in South Africa." *American Economic Review*, 86:2, pp. 335-39.

Mwabu, G. and T. P. Schultz (2000) "Wage Premiums for Education and Location of South African Workers by Gender and Race", *Economic Development and Cultural Change*, 4:2, pp. 307-334.

Mincer, J. (1958) "Investment in Human Capital and Personal Income Distribution" *The Journal of Political Economy*, 66:4, pp. 281-302.

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

Nattrass, N. (2000) "The Debate About Unemployment in the 1990s", *Studies in Economics and Econometrics*, 24:3, pp. 129-142.

Posel, D. (1999) "Migration, Poverty Traps & Development: A Tale of Two Villages", University of Natal, Durban, Discussion Paper Series DP-3

Psachararopoulos, G (1973) "Returns to education. A further International Update", Elsevier Scientific Publishing Company. Amsterdam.

Psachararopoulos, G (1985) "Returns to Eucation. A further International Update", The Journal of Human Resources, 20:4 pp.583-604.

Psachararopoulos, G (1994) "Returns to Investment in Education", *World Development*, 22:9, pp. 1325-43.

Psachararopoulos, G. and H. Patrinos (2002) "Returns to Investment in Education: A Further Update", *World Bank Policy Research Working Paper*, Number 2881.

Romer, Paul M. (1986) "Increasing Returns and Long-run Growth", *The Journal of Political Economy*, 94:5, pp1002-1037.

Rosen, Sherwin (1992) "Distinguished Fellow: Mincering Labour Economics", *Journal of Economic Perspectives*, 6:2, pp. 157-70.

Rospabe, S (2001) "An Empirical Evaluation of Gender Discrimination in Employment, Occupation Attainment and Wage in South Africa in the late 1990s." Paper presented at DPRU/ FES Conference on Labour Markets and Poverty in South Africa: Misty Hills, Johannesburg.

Schultz, T.P (1988) "Education Investment and Return" in H. Chenery and T.N. Srinivasan, eds., *Handbook of Development Economics*, Vol 1 Amsterdam: Elsevier Science Publishers.

Siphambe, H.P. (2000) "Rates of return to education in Botswana", *Economics of Education Review, 19*, pp291-300.

Skyt-Nielsen, H. and N. Westergård-Nielsen (1998) "Returns to Schooling in LDCs: New Evidence from Zambia", Working Paper No. *98-10*, Centre for Labour Market and Social Research, University of Aarhuis and the Aarhuis School of Business.

Stata Reference Manual Release 7 Volume 2 H-P (2001) College Station, Texas: Stata Press.

Tamura, R. (1991) "Income Convergence in an Endogenous Growth Model", *The Journal of Political Economy*, 99:3, pp. 522-540.

Teal, F. (2001) "Education, Incomes, Poverty and Inequality in Ghana in the 1990s", Working Paper No. 2001-21, Centre for the Study of African Economies, Oxford University.

Whaba, J. (2000) "Returns to Education and Regional Earnings Differentials in Egypt", Discussion Papers in Economics and Econometrics, University of South Hampton, United Kingdom.

Wittenberg, M. (1999) "Job Search and Household Structure in an Era of Mass Unemployment: a semi-parametric analysis of the South African Labour Market." Working Paper No. 3, ERSA, University of Witwatersrand. South Africa.

Appendix A: Estimates of Mincerian Returns to Schooling in South African Studies

Study	Coverage of earnings function	Regression type	Dependent variable	Controls	Education terms				
Mwabu and Schultz (2000) (PSLSD 1993) Source:Table 2 p214	Wage earners 16-65 Separate equations for gender, race, location(not included here)	Heckman and OLS. Reported results for OLS only as little difference	log gross hourly wage	rural dummy, experience, experience^2	Splines	African Male	African Female	White Male	White Female
					Primary	0.084 *	0.062 *	-0.012	-0.034
					Secondary	0.158 *	0.249 *	0.084 *	0.052 *
					Tertiary	0.294 *	0.396 *	0.151 *	0.139 *
					n	9325	10473	1447	1517
Rospabe (2001) (OHS 1999)	Formal & informal 16-65	multinomial logit	unemployed, employed, self-employed	race dummies, age, age^2, no of kids, urban, married, family members' employment status, household head, distance from phone, province	Splines	Male employed	Male self-employed	Female employed	Female self-employed
					Primary	-0.022 *	0.031	-0.031 *	-0.052 *
					Secondary	0.039 *	0.032	0.086 *	-0.051 *
					Tertiary	0.446 *	0.422 *	0.764 *	0.636 *
					n	19920		18913	
		interval regression (generalized tobit)	log groass hourly earnings	race, experience, experience^2, tenure, tenure^2, union, occupation, industrial sector, province, urban, married, formal sector	Splines	Males	Females		
					Primary	0.027 *	0.030 *		
					Secondary	0.091 *	0.111 *		
					Tertiary	0.176 *	0.153 *		
					n	9913	7651		
Bhorat (2000) (OHS, 1995)	Formal & informal 16-64 Separate equations for race	OLS	log monthly wage	gender, province ,sector, union, experience, experience^2, urban, log hours worked per month	Splines	African skilled	White skilled	African semi-skill	White semi-skill
					Primary	0.037 *	0.010	0.039 *	-0.074
					Secondary	0.122 *	-0.004	0.128 *	0.137 *
					Tertiary	0.159 *	0.256 *	0.017	-0.031
					n	2663	2536	7396	3179
Michaud and Vencatachellum (2001) (PSLSD 1993)	Wage earners	OLS	log gross hourly wage	experience, experience^2, wealth proxy, skilled/semi-skilled dummy, manufacturing/ tertiary/ professional sector dummy, urban, province, union dummy	Splines	African male	African female		
					Primary	0.046 *	0.013		
					Secondary	0.116 *	0.073 *		
					Tertiary	0.114 *	0.055		
					n	2361	1525		
		Modified Heckman two step to estimate the wage equation reported here	Stage 1: (Bivariate probit model with selection to account for endogenous union membership and labour market participation)	experience, experience^2, wealth proxy, skilled/semi-skilled dummy, manufacturing/ tertiary/ professional sector dummy, urban, province, union dummy	Splines	African male non-unionised	Africa male unionized	African female non-unionised	African female unionized
					Primary	0.042 *	-0.007	0.013	-0.066 *
					Secondary	0.114 *	0.055 *	0.052 *	0.092 *
					Tertiary	0.085 *	0.182 *	0.035	0.136 *
					n	1571	791	1208	317

Study	Coverage of earnings function	Regression type	Dependent variable	Controls	Education terms				
Bhorat & Leibbrandt	Africans only	Probit	participation	age dummies, urban,					
(2001)	16-65			household structure variables,	Splines	Broad UE		Narrow UE	
(OHS, 1995)		Marginal effects		household income		African male	African female	African male	African female
		At the mean			Primary	0.003 *	0.003	0.004 *	0.003
					Secondary	0.005 *	0.052 *	0.016 *	0.057 *
					Tertiary	0.003	-0.023	0.016 *	0.011 *
					n	15658	19548	15658	19548
		Probit	employment	age dummies, urban, province	Splines	Broad UE		Narrow UE	
		Marginal effects				African male	African female	African male	African female
		at the mean			Primary	-0.012 *	-0.004 *	-0.009 *	-0.009 *
					Secondary	0.010	0.036 *	0.000	-0.021 *
					Tertiary	0.047 *	0.142 *	0.036 *	0.153 *
					n	14203	12810	11931	9426
	Formal & informal	Heckman 2 step,	log gross	urban, province, industrial	Splines	Broad UE		Narrow UE	
		ML	monthly wage	sector, occupation,		African male	African female	African male	African female
				experience, experience^2, log	Primary	0.035 *	0.049 *	0.036 *	0.051 *
	(excluding those			hours worked per month	Secondary	0.109 *	0.082 *	0.108 *	0.093 *
	identified as both)				Tertiary	0.037	0.023	0.031	0.032 *
					n	14124	12723	11886	9393
Lam (1999)	(doesn't state whether	OLS	log gross	White dummy, age, age^2	Dummies	Males			
(OHS, 1995)	formal &		monthly		1-3 yrs	-0.010			
	informal)		earnings		4	0.090 *			
	Males, all races				5	0.150 *			
	30-49				6	0.269 *			
					7	0.397 *			
					8	0.571 *			
					9	0.733 *			
					10	0.968 *			
					11	1.041 *			
					12	1.484 *			
					>=15	1.970 *			
					n	10867			
Moll (1996)	Africans, urban,male,	OLS	log of gross	experience, experience^2,	Dummies	African males			
(CSS/HSRC1990)	non agricultural		annual cash	married, region dummies	0	-0.140 *			
Central Statistical	employees		income		1-3 yrs	-0.170 *			
Service and Human	18-59				4	-0.110 *			
Sciences Research					5	-0.075 *			
Council Survey					6	-0.088 *			
				Moll is the only author that	7_base	0			
				assigns a dummy to 12 years	8	0.068 *			
				of schooling as well as to a	9	0.120 *			
				diploma or degree. The	10	0.210 *			
				coefficient for higher	11	0.340 *			
				education in this case is	12	0.540 *			
				interpreted as additional to	Diploma	0.330 *			
				the schooling increases	Degree	0.480 *			
					n	2855			

Study	Coverage of earnings function	Regression type	Dependent variable	Controls	Education terms				
Hofmeyer (2001) (OHS 1999)	African males Formal & informal 16 and over (reporting results for most relevant sectors)	multinomial logit coefficients to be interpreted as odds ratios	Inactive, unemployed, informal sector, unregulated formal employment, formal employment non-union member, formal employment union member*	urban, rural, single, not household head age,age^2, grant support, household size, have young children	Dummies	Formal non-union	Formal union		
					primary	-0.418 *	-0.131		
					std8	-0.509 *	-0.357 *		
					matric	0.683 *	1.048 *		
					dip (no mat)	0.161	1.088 *		
					dip (with mat)	0.973 *	1.963 *		
					degree	1.001 *	1.681 *		
		weighted least squares (wage equation)	log gross hourly wage	urban, rural, industrial sector, occupation, single, not household head, experience, experience^2, job duration, (job duration)^2	Dummies	Formal non-union	Formal union		
					primary	0.175 *	0.140 *		
					std8	0.326 *	0.247 *		
					matric	0.507 *	0.470 *		
					dip (no mat)	0.486 *	0.623 *		
					dip (with mat)	0.957 *	0.841 *		
					degree	1.297 *	1.006 *		
					n	3039	2720		
Hosking (2001) (1996 Population Census)	Not stated	OLS	log earnings (gives no further detail)	experience, experience^2		African	Coloured	Indian	White
					Primary	0.38 *	0.52 *	0.1 0 *	-0.43 *
					Secondary	1.58 *	1.79 *	1.01 *	0.56 *
					Tertiary	2.27 *	2.33 *	1.58 *	0.98 *
					n	446938	88442	31083	154695
						All Races			
					Primary	0.40 *			
					Secondary	1.95 *			
					Tertiary	2.52 *			
					n	727098			
Kingdon & Knight (1999) (PSLSD 1993)	wage earners 16-64 (doesn't state whether formal & informal)	OLS	log gross hourly wage	experience, experience^2, race, gender, union, urban, province, married, occupation, public sector, tar road dummies, broad U^e rate per residential cluster, (broad U^e rate)^2, dissatisfaction, homeland		All races & genders			
					Years of education	0.004			
					(Years of education)^2	0.004 *			
					n	6498			
Erichsen & Wakeford (2001) (OHS, 1995, PSLSD 1993)	Regular wage earning employees 16-64 Excludes self-employed	OLS	log gross monthly wage	experience, experience^2,rural, union, industrial sector, occupation, province		1993 male	1995 male	1993 female	1995 female
					Years of education	-0.004	-0.005	-0.01	0.026 *
					(Years of education)^2	0.007 *	0.007 *	0.005 *	0.004 *
					n	2223	11727	1570	6396

Note: * Indicates significance up to the 10% level

Appendix B: Issues Concerning the Definition of Unemployment, Employment and Earnings

The aim of the empirical analysis was to test the same earnings function specification across a number of datasets and thereby to verify whether increasing returns are a robust feature of the data. Working with a number of datasets required a careful study of each survey's representativeness. It also required definitions of variables to be precise and consistent across datasets. In this Appendix we discuss our choice of definitions used and how we tested for the representativeness of the sample surveys.

In the econometric exercise we wished to explore the total effect of education on individuals' economic outcomes. We therefore include the unemployed in the analysis in order to account of how education impacts on both the probability of finding employment and on ones earnings given one is employed. As the unemployed are included, it was necessary to define the economically active population and therefore choose whether to use the broad or narrow definitions of unemployment. A growing literature on the extent of 'discouraged workers' and the greater suitability of using the expanded definition when examining the jobless in a South African context informed our choice here (see Kingdon and & Knight (1999), (2000); Nattrass (2000); Wittenberg(1999); Dinkelman & Pirouz (2002)).

Although we use a broad classification for the unemployed, we use a 'narrow' definition for the employed. Income earners are restricted to formal sector full time wage employees from age 15. The main reason for using this restricted group is that full time wage employment should represent the least variable component in earnings. Self-employed, part-time and casual workers will have highly variable earnings that will often be subject to a large degree of seasonality. Inclusion of these groups will diminish the reliability of the coefficients across surveys as the nature of these data are inherently noisy. Furthermore, there is not adequate income and expenditure information for the self-employed in the post 1995 OHSs and the LFS, 2000. In addition to the above, definitions of part-time, casual, seasonal and temporary work also differ across surveys.

All individuals who were working in some form but did not fit our definition of employment are excluded from the sample of interest. The economically active population then includes all formal sector full time wage employees and those classified as unemployed according to the broad definition of unemployment thereby including discouraged workerseekers, that is, those who indicate they would like to work even if they are not actively seeking employment.

Once the appropriate definitions of the relevant variables had been established, it was necessary to deal with two particular problems concerning the income variable. The first problem is that of missing values. Owing to the private nature of individual earnings, this variable reflected the highest level of non-response. Following Hertz (2001) these respondents are simply excluded from our estimates. [27]

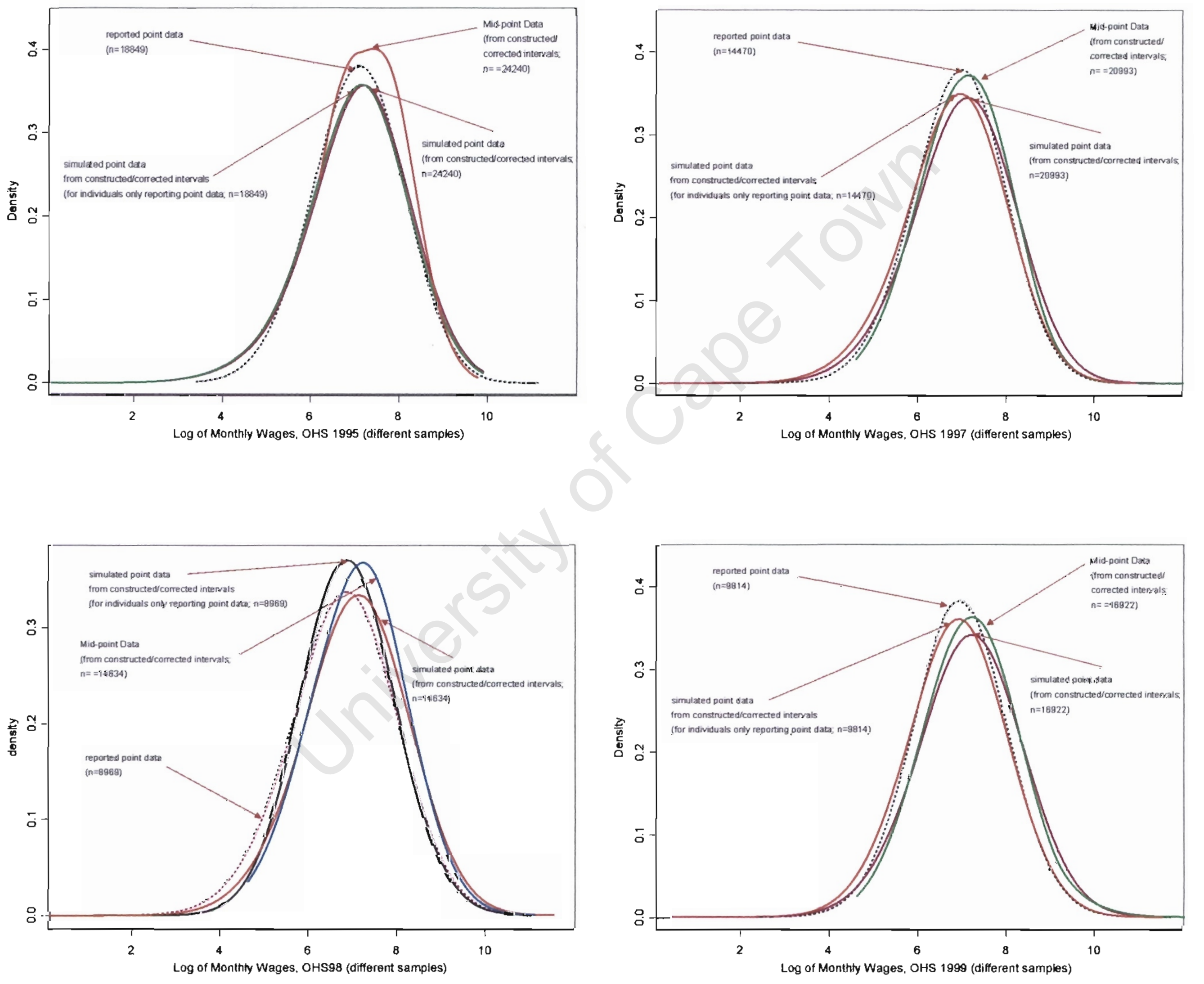
The second problem arises in the Statistics South Africa surveys in which respondents were given the option of reporting their actual earnings or alternatively, earnings categories. Percentage-wise, the number who reported their actual or point income in 1997, 1998 and 1999 is relatively low. It was therefore necessary to ascertain whether reporting of point as opposed to interval data was random. If this was the case, we could exclude those with interval data from our estimations. If, however, there was some systematic process at work that resulted in a bias in the distribution of point income reported as opposed to interval, one could not use only the point data as assumptions of log normality of the distributions would be violated. Moreover, ordinary least squares regressions cannot be run on midpoints of interval data as the data is count, not continuous. We explain and demonstrate below how we tested for randomness of reported point income in each of the OHS datasets.

The samples of interest are all those classified as employed according to our definition, who have reported income, either as an actual amount or as falling into a specified earnings bracket. The first step was to create an income category variable that reflected the appropriate income interval for all individuals under consideration. Step two involved simulating point data for each earnings category assuming random sampling with replacement and a uniform distribution within the categories. The next step was to construct a density plot for the simulated point data so that we could visualize the income distribution of the entire sample. We then generated a new variable that reflected intervals for those cases where only point data was reported and repeat the simulation process on this variable. Finally density plots using midpoint data for the full set of income categories as well as for just those who reported point data were also constructed. *The*

[27] Using the PSLD 1993 Hertz(2001) compares the distribution of education across those who have not reported an income with those who have reported an income and finds them extremely similar. This infers that the incidence of earnings non-reporting is independent of a respondent's schooling level. The random nature of non-reporting then means that the missing values can be excluded without biasing our results.

aim of constructing the various plots was to develop a representative distribution and then to compare this with the other various earnings distribution plots to ascertain if the point data we have appears random and can be used in estimation techniques to produce unbiased estimates. Figure 4 in which the histograms are kernel density estimates integrated to 1, shows the results.

Figure 4: Testing the representativeness of differently reported income measures in the October Household Surveys



Source: Keswell and Poswell (2002)
Notes:
The figures are density plots of reported and simulated data used to test the randomness of reporting income either as an actual figure or according to an income category. All histograms are kernel density estimates integrated to one.

We start by comparing the density plots for those who reported point income, with the simulated data from constructed income categories for this exact group of respondents, to examine how well our simulation rule replicates the true distribution. We see that for 1995 and 1997 the density of the actual point data exceeds that of the simulated point data indicating slightly more variation in the log of wages for the simulated case. The reverse is true, however, for 1998 and 1999 with the reported data having the flatter distribution. These simulated plots, however, lie fairly closely to the reported data plots indicating a reasonably representative simulation.

Across surveys, the density plot that diverges most from the plot of actual reported data is the graph constructed from interval midpoints for all individuals. This is most apparent for the 1995 OHS where we see the midpoint line is rather arbitrary and has quite obviously over-sampled wealthier individuals in its assigning observations to the middle of each relevant category. The reason the OHS, 1995 result is so pronounced is rooted in the questionnaire design for this year. The choice of income categories in 1995 differs from the other years in that the lowest category was a rather large R1-R999 (per week, month or year) whereas in subsequent years this income range was broken down into 3 smaller categories which allows for better approximation of the actual data, especially in a country such as South Africa where such a large proportion of the population are low earners. It can be seen quite clearly from our graphs that using midpoint data will not yield unbiased results.

Finally we compare the density plot derived from simulated point data across the income intervals covering all reported earnings with the plot for which only point data is reported. We see for 1995, 1998 and 1999 that the plot for point data lies slightly to the left of that for the entire sample. This would seem to indicate that there is slightly lower response of actual income information for those in the higher income brackets. We can also see that for 1995, 1997 and 1999 the wage dispersion is not as great for the reported data as it is for the entire simulated sample and that in the reported point data we appear to have undersampling of the poor and oversampling of the middle classes.

The differences for 1995, 1997 and 1998 however, appear small and we are therefore satisfied that for the purposes of our investigation that the point data is (*surprisingly*) representative of the entire sample. From this evidence we then decide to reduce our samples in the OHS surveys to include only those who have reported point income. We acknowledge that this will raise the sample "unemployment rate" above the actual unemployment rate for each year. The randomness of the point income reported, however, should mean that no large bias is introduced. OHS, 1999 reveals the most severe difference between the reported point and interval income data and seemingly the highest probability that estimates based on this data will be biased. This is a key reason for excluding the survey and results from our analysis here. We also then exclude OHS, 1998 partially as it too has a large percentage of non-response but more importantly as LFS, 2000 is actually based on the same major sample but has better captured point earnings reported.

Appendix C: Issues concerning censoring bias

Simple Mincerian returns to education are generally calculated using ordinary least squares regressions run for all those with positive earnings in the relevant sample. This, however, can be problematic. Random samples of a population of interest are likely to include both earners and non-earners. Excluding non-earners from the regression means that the sample is no longer randomly selected and OLS results will not be truly representative of the population. (Breen (1996: 2)) Furthermore, the results will be biased for statistical reasons that are elaborated on below. Essentially we have a sample selection problem in that we have the right hand side variables for all individuals sampled, but either a zero for those who are unemployed or positive income for those who are employed. That is, one can only earn an income, given one is employed. Our dependent variable is therefore truncated from below at zero.

In a standard OLS regression of the form

$$y_i = x_i'\beta + \mu_i \tag{1}$$

the estimates of β are the BLUE if the assumptions of $E(\mu) = 0$ and $E(\mu_i\mu_j) = 0$ if $i \neq j$ hold or equivalently if $\mu_i \sim IN(0, \sigma^2)$. If either of these assumptions is violated, then estimates of β will be biased.

When using OLS as an estimation technique, one has the choice of running the regression on the sub-sample of the employed only or of running it on the full sample (that is, including both the employed and the unemployed).

If the first option is chosen and only those who report positive earnings are included (thereby creating a truncated sample) y_i is observed only if $y_i > 0$. Specifically,

$$y_i = x_i'\beta + \mu_i \qquad \text{only if } y_i > 0 \tag{1a}$$

Estimating such a model requires one to take conditional expectations according to the formula for the expected value of a random variable y_i conditional on x_i and censored at c (in this case c=0). The formula is shown in equation 2 below.

$$E(y_i \mid x_i) = pr(y_i > 0 \mid x_i)E(y_i \mid y_i > 0, x_i) \tag{2}$$

Now $y_i > 0$ implies that

$$\beta x_i + \mu_i > 0 \quad \text{or} \quad \mu_i > -\beta x_i \tag{3}$$

so to estimate the first part of the equation 2, or $pr(y_i > 0 \mid x_i)$, we must find

$$pr(u_i > -xi'\beta) = pr(ui < xi'\beta) \tag{4}$$

The solution is to find the probability of the random variable μ_i that is assumed normally distributed with mean zero $\{E(\mu_i) = 0\}$and constant variance , σ^2. The answer can then be found by applying the formula for the cumulative standard normal distribution function given in equation 5 below.

$$\Phi_i = \Phi(\frac{x_i'\beta}{\sigma}) = \int_{-\infty}^{x_i'\beta/\sigma} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{5}$$

To estimate the second part of the equation we need to find $E(y_i \mid y_i > 0, x_i)$:

Now

$$E(y_i \mid y_i > 0, x_i) = x_i'\beta + E(\mu_i \mid \mu_i > -x_i'\beta) \tag{6}$$

To estimate such an equation using normal OLS, it is required that

$$E(\mu_i \mid \mu_i > -x_i'\beta) = 0 \tag{7}$$

However, it is already assumed that the unconditional expected value of μ_i is zero $\{E(\mu_i) = 0\}$ so clearly the conditional expected value of μ_i $\{E(\mu_i \mid \mu_i > - \beta x_i)\}$ cannot also equal zero. It is apparent that in this case two assumptions for unbiased OLS estimate of β are violated, namely that the expected value of μ is zero and that μ and x are not correlated. (Maddala (1999:2); Breen (1996:26))

Developing the model more fully shows why running OLS on all the observations will also yield biased results:

To get an unbiased estimate of $E(\mu_i \mid \mu_i > -x_i'\beta)$, it is necessary to use the statistical result that gives the expected value of a truncated normally distributed random variable as shown below.

$$\phi_i \equiv \phi\left(\frac{x_i'\beta}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{(-x_i'\beta)^2 / 2\sigma^2} \tag{8}$$

where ϕ_i is the standard normal density function evaluated at $x_i'\beta/\sigma$

The solution to equation 7 then includes the evaluation of the standard normal distribution function and the standard normal density function at $x_i'\beta/\sigma$

$$E(u_i \mid u_i > -x_i'\beta) = \sigma \frac{\phi_i}{\Phi_i} \tag{9}$$

Now to estimate the complete model

$$E(y_i \mid x_i) = pr(y_i > 0 \mid x_i) E(y_i \mid y_i > 0, x_i)$$

equations 5 and 9 are combined to get

$$E(y_i \mid x_i) = \Phi_i \left[x_i'\beta + \sigma \frac{\phi_i}{\Phi_i} \right] \tag{10}$$

From equation 10 it can be seen why running OLS on all observations will also lead to biased results. This is because in order to get an unbiased estimate of β it is necessary that $\Phi_i = 1$ and therefore,e that $\phi_i = 0$ (ie none of the observations in the sample be censored). But, the sample *is* censored so this is clearly problematic and the OLS estimates will be biased. (Breen (1996:25))

Appendix D: Further Estimates

Table 10: Marginal Effects for OLS and Tobit Estimators (at the mean) for Specifications 2 to 9

	OLS93	Tobit93	OLS95	Tobit95	OLS97	Tobit97	OLS00	Tobit00
Specification 2								
Constant	-4.524 *	-8.573 *	-4.591 *	-8.390 *	-4.160 *	-8.449 *	-6.291 *	-11.727 *
	(0.316)	(0.376)	(0.178)	(0.218)	(0.202)	(0.236)	(0.185)	(0.220)
Age	0.340 *	0.409 *	0.336 *	0.411 *	0.292 *	0.367 *	0.393 *	0.513 *
	(0.018)	(0.021)	(0.009)	(0.011)	(0.011)	(0.013)	(0.010)	(0.011)
Age squared	-0.003 *	-0.004 *	-0.003 *	-0.004 *	-0.002 *	-0.003 *	-0.003 *	-0.005 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	-0.273 *	-0.270 *	-0.210 *	-0.218 *	-0.177 *	-0.179 *	-0.206 *	-0.192 *
	(0.029)	(0.032)	(0.015)	(0.018)	(0.016)	(0.017)	(0.018)	(0.019)
(Education in years)^2	0.039 *	0.037 *	0.034 *	0.034 *	0.026 *	0.025 *	0.028 *	0.025 *
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
sigma	3.161		3.181		3.282		3.287	
R^2	0.196		0.206		0.136		0.207	
Specification 3								
Constant	-4.731 *	-8.731 *	-4.704 *	-8.475 *	-4.238 *	-8.481 *	-6.490 *	-11.778 *
	(0.318)	(0.378)	(0.178)	(0.218)	(0.202)	(0.236)	(0.184)	(0.219)
Age	0.340 *	0.412 *	0.332 *	0.407 *	0.286 *	0.361 *	0.384 *	0.504 *
	(0.018)	(0.021)	(0.009)	(0.011)	(0.011)	(0.013)	(0.010)	(0.011)
Age squared	-0.003 *	-0.004 *	-0.003 *	-0.004 *	-0.002 *	-0.003 *	-0.003 *	-0.005 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	0.098	-0.064	0.122 *	0.065 **	0.218 *	0.127 *	0.300 *	0.146 *
	(0.062)	(0.071)	(0.033)	(0.039)	(0.034)	(0.038)	(0.036)	(0.039)
(Education in years)^2	-0.035 *	-0.003	-0.029 *	-0.019 *	-0.052 *	-0.035 *	-0.063 *	-0.035 *
	(0.011)	(0.013)	(0.006)	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)
(Education in years)^3	0.004 *	0.002 *	0.003 *	0.003 *	0.004 *	0.003 *	0.004 *	0.003 *
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sigma	3.152		3.174		3.273		3.273	
R^2	0.201		0.209		0.140		0.214	
Specification 4								
Constant	-3.498 *	-7.280 *	-3.043 *	-6.501 *	-1.804 *	-5.514 *	-3.600 *	-8.305 *
	(0.292)	(0.356)	(0.165)	(0.208)	(0.178)	(0.217)	(0.163)	(0.202)
Age	0.266 *	0.334 *	0.243 *	0.308 *	0.167 *	0.225 *	0.242 *	0.344 *
	(0.016)	(0.020)	(0.009)	(0.011)	(0.010)	(0.012)	(0.008)	(0.010)
Age squared	-0.003 *	-0.003 *	-0.002 *	-0.003 *	-0.001 *	-0.002 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	0.084	-0.074	0.123 *	0.074 **	0.127 *	0.066 **	0.180 *	0.080 **
	(0.057)	(0.067)	(0.031)	(0.037)	(0.030)	(0.034)	(0.031)	(0.035)
(Education in years)^2	-0.040 *	-0.011	-0.037 *	-0.030 *	-0.039 *	-0.030 *	-0.045 *	-0.028 *
	(0.010)	(0.012)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)
(Education in years)^3	0.004 *	0.002 *	0.003 *	0.003 *	0.003 *	0.002 *	0.003 *	0.002 *
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Union	3.726 *	3.422 *	3.241 *	3.310 *	4.420 *	3.823 *	4.474 *	3.675 *
	(0.091)	(0.099)	(0.043)	(0.051)	(0.045)	(0.050)	(0.046)	(0.050)
sigma	2.879		2.927		2.860		2.847	
R^2	0.333		0.327		0.343		0.405	

	OLS93	Tobit93	OLS95	Tobit95	OLS97	Tobit97	OLS00	Tobit00
Specification 5								
Constant	-3.249 *	-7.004 *	-2.399 *	-5.714 *	-1.516 *	-5.165 *	-3.302 *	-8.005 *
	(0.291)	(0.355)	(0.161)	(0.205)	(0.176)	(0.214)	(0.162)	(0.201)
Age	0.272 *	0.338 *	0.247 *	0.312 *	0.179 *	0.237 *	0.248 *	0.352 *
	(0.016)	(0.020)	(0.008)	(0.011)	(0.010)	(0.012)	(0.008)	(0.010)
Age Squared	-0.003 *	-0.003 *	-0.002 *	-0.003 *	-0.001 *	-0.002 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.649 *	-0.686 *	-1.459 *	-1.817 *	-0.939 *	-1.035 *	-0.758 *	-0.849 *
	(0.063)	(0.074)	(0.033)	(0.041)	(0.032)	(0.038)	(0.033)	(0.039)
Education in years	0.061	-0.096	0.080 *	0.020	0.109 *	0.048 **	0.167 *	0.069 **
	(0.056)	(0.066)	(0.030)	(0.037)	(0.029)	(0.034)	(0.031)	(0.035)
(Education in years)^2	-0.035 *	-0.006	-0.027 *	-0.017 *	-0.035 *	-0.026 *	-0.043 *	-0.026 *
	(0.010)	(0.012)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)
(Education in years)^3	0.004 *	0.002 *	0.003 *	0.002 *	0.003 *	0.002 *	0.003 *	0.002 *
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Union	3.623 *	3.316 *	3.021 *	3.068 *	4.247 *	3.637 *	4.327 *	3.509 *
	(0.091)	(0.099)	(0.042)	(0.050)	(0.045)	(0.050)	(0.046)	(0.050)
sigma	2.861		2.839		2.823		2.823	
R^2	0.341		0.367		0.361		0.416	
Specification 6								
Constant	0.459	-3.571 *	0.237	-3.025 *	0.589 *	-3.437 *	-0.457 *	-5.497 *
	(0.300)	(0.368)	(0.164)	(0.211)	(0.189)	(0.227)	(0.173)	(0.213)
Age	0.254 *	0.320 *	0.253 *	0.321 *	0.206 *	0.269 *	0.250 *	0.353 *
	(0.015)	(0.019)	(0.008)	(0.010)	(0.009)	(0.011)	(0.008)	(0.010)
Age Squared	-0.003 *	-0.003 *	-0.002 *	-0.003 *	-0.002 *	-0.003 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.630 *	-0.683 *	-1.385 *	-1.746 *	-0.859 *	-0.952 *	-0.712 *	-0.801 *
	(0.059)	(0.071)	(0.031)	(0.039)	(0.031)	(0.037)	(0.032)	(0.038)
African	-3.406 *	-3.029 *	-2.710 *	-2.782 *	-2.707 *	-2.427 *	-2.870 *	-2.495 *
	(0.111)	(0.124)	(0.054)	(0.064)	(0.081)	(0.088)	(0.071)	(0.077)
Coloured	-1.918 *	-1.419 *	-1.341 *	-1.119 *	-0.961 *	-0.506 *	-1.279 *	-0.744 *
	(0.139)	(0.156)	(0.064)	(0.077)	(0.089)	(0.097)	(0.082)	(0.089)
Asian	-1.388 *	-1.047 *	-0.583 *	-0.460 *	-0.846 *	-0.561 *	-0.848 *	-0.572 *
	(0.181)	(0.199)	(0.091)	(0.108)	(0.137)	(0.148)	(0.125)	(0.134)
Education in years	0.063	-0.052	0.095 *	0.052	0.069 **	0.013	0.118 *	0.038
	(0.053)	(0.064)	(0.028)	(0.035)	(0.028)	(0.033)	(0.029)	(0.034)
(Education in years)^2	-0.030 *	-0.011	-0.034 *	-0.029 *	-0.030 *	-0.022 *	-0.036 *	-0.022 *
	(0.010)	(0.011)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.005)
(Education in years)^3	0.003 *	0.002 *	0.003 *	0.003 *	0.002 *	0.002 *	0.002 *	0.002 *
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Union	3.441 *	3.149 *	3.107 *	3.201 *	4.144 *	3.543 *	4.341 *	3.546 *
	(0.087)	(0.096)	(0.040)	(0.048)	(0.043)	(0.048)	(0.044)	(0.048)
sigma	2.689		2.697		2.711		2.703	
R^2	0.418		0.429		0.410		0.464	

	OLS93	Tobit93	OLS95	Tobit95	OLS97	Tobit97	OLS00	Tobit00
Specification 7								
Constant	1.393 *	-2.671 *	0.179	-3.469 *	0.962 *	-3.123 *	-0.387	-5.595 *
	(0.415)	(0.515)	(0.242)	(0.318)	(0.260)	(0.317)	(0.286)	(0.356)
Age	0.218 *	0.288 *	0.249 *	0.330 *	0.185 *	0.252 *	0.244 *	0.351 *
	(0.018)	(0.022)	(0.010)	(0.013)	(0.011)	(0.013)	(0.010)	(0.013)
Age Squared	-0.002 *	-0.003 *	-0.002 *	-0.003 *	-0.002 *	-0.002 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.622 *	-0.687 *	-1.388 *	-1.753 *	-0.855 *	-0.949 *	-0.711 *	-0.800 *
	(0.059)	(0.071)	(0.031)	(0.039)	(0.031)	(0.037)	(0.032)	(0.038)
African	-3.317 *	-3.007 *	-2.661 *	-2.771 *	-2.588 *	-2.358 *	-2.798 *	-2.479 *
	(0.112)	(0.125)	(0.054)	(0.065)	(0.081)	(0.088)	(0.072)	(0.077)
Coloured	-1.836 *	-1.401 *	-1.290 *	-1.106 *	-0.844 *	-0.437 *	-1.210 *	-0.730 *
	(0.140)	(0.157)	(0.065)	(0.078)	(0.090)	(0.097)	(0.083)	(0.089)
Asian	-1.322 *	-1.060 *	-0.542 *	-0.462 *	-0.734 *	-0.499 *	-0.802 *	-0.577 *
	(0.182)	(0.200)	(0.091)	(0.109)	(0.137)	(0.149)	(0.125)	(0.134)
Education in years	0.176	-0.063	0.831 *	0.927 *	0.574 *	0.636 *	0.482 *	0.445 *
	(0.171)	(0.212)	(0.097)	(0.126)	(0.102)	(0.122)	(0.106)	(0.130)
(Education in years)^2	-0.081 *	-0.050	-0.186 *	-0.211 *	-0.133 *	-0.155 *	-0.105 *	-0.106 *
	(0.031)	(0.038)	(0.017)	(0.021)	(0.018)	(0.022)	(0.017)	(0.020)
(Education in years)^3	0.005 *	0.005 **	0.010 *	0.012 *	0.007 *	0.008 *	0.005 *	0.006 *
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age* Education	-0.005	0.000	-0.020 *	-0.023 *	-0.015 *	-0.018 *	-0.011 *	-0.010 *
	(0.005)	(0.006)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Age* Education^2	0.002 **	0.001	0.004 *	0.005 *	0.003 *	0.004 *	0.002 *	0.002 *
	(0.001)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Age* Education^3	0.000 **	0.000 **	0.000 *	0.000 *	0.000 *	0.000 *	0.000 *	0.000 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Union	3.424 *	3.137 *	3.095 *	3.189 *	4.109 *	3.523 *	4.322 *	3.543 *
	(0.087)	(0.096)	(0.040)	(0.048)	(0.043)	(0.048)	(0.044)	(0.048)
sigma	2.686		2.692			2.706		2.701
R^2	0.420		0.431		0.413		0.465	
Specification 8								
Constant	-4.224 *	-8.047 *	-3.393 *	-6.827 *	-1.973 *	-5.652 *	-4.053 *	-8.788 *
	(0.294)	(0.357)	(0.161)	(0.203)	(0.175)	(0.213)	(0.158)	(0.197)
Age	0.271 *	0.334 *	0.251 *	0.314 *	0.181 *	0.238 *	0.255 *	0.358 *
	(0.017)	(0.020)	(0.009)	(0.011)	(0.010)	(0.012)	(0.008)	(0.010)
Age Squared	-0.003 *	-0.003 *	-0.002 *	-0.003 *	-0.001 *	-0.002 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.678 *	-0.709 *	-1.468 *	-1.824 *	-0.940 *	-1.034 *	-0.745 *	-0.834 *
	(0.064)	(0.076)	(0.033)	(0.041)	(0.033)	(0.039)	(0.033)	(0.039)
Education in years	0.206 *	0.196 *	0.220 *	0.231 *	0.087 *	0.070 *	0.118 *	0.103 *
	(0.008)	(0.010)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)
Union	3.606 *	3.284 *	3.042 *	3.069 *	4.331 *	3.704 *	4.431 *	3.585 *
	(0.093)	(0.101)	(0.043)	(0.050)	(0.045)	(0.050)	(0.046)	(0.050)
sigma	2.944		2.898		2.845		2.853	
R^2	0.303		0.341		0.350		0.403	

	OLS93	Tobit93	OLS95	Tobit95	OLS97	Tobit97	OLS00	Tobit00
Specification 9								
Constant	0.484	-3.635 *	-0.085	-3.460 *	0.525 *	-3.557 *	-0.744 *	-5.843 *
	(0.301)	(0.370)	(0.165)	(0.211)	(0.189)	(0.227)	(0.171)	(0.211)
Age	0.253 *	0.316 *	0.256 *	0.323 *	0.207 *	0.270 *	0.256 *	0.357 *
	(0.015)	(0.019)	(0.008)	(0.010)	(0.009)	(0.011)	(0.008)	(0.010)
Age Squared	-0.002 *	-0.003 *	-0.002 *	-0.003 *	-0.002 *	-0.003 *	-0.002 *	-0.003 *
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.641 *	-0.690 *	-1.389 *	-1.748 *	-0.858 *	-0.947 *	-0.700 *	-0.786 *
	(0.060)	(0.072)	(0.031)	(0.040)	(0.031)	(0.037)	(0.032)	(0.038)
African	-3.868 *	-3.456 *	-3.015 *	-3.077 *	-2.987 *	-2.684 *	-3.106 *	-2.695 *
	(0.106)	(0.119)	(0.053)	(0.064)	(0.080)	(0.087)	(0.071)	(0.076)
Coloured	-2.457 *	-1.922 *	-1.725 *	-1.501 *	-1.294 *	-0.813 *	-1.550 *	-0.975 *
	(0.135)	(0.150)	(0.064)	(0.076)	(0.088)	(0.095)	(0.082)	(0.088)
Asian	-1.635 *	-1.268 *	-0.777 *	-0.654 *	-1.051 *	-0.750 *	-0.979 *	-0.682 *
	(0.182)	(0.199)	(0.092)	(0.109)	(0.137)	(0.149)	(0.126)	(0.134)
Education in years	0.088 *	0.075 *	0.119 *	0.120 *	0.046 *	0.028 *	0.062 *	0.046 *
	(0.008)	(0.010)	(0.004)	(0.006)	(0.004)	(0.005)	(0.005)	(0.005)
Union	3.442 *	3.144 *	3.132 *	3.212 *	4.217 *	3.605 *	4.426 *	3.611 *
	(0.088)	(0.097)	(0.041)	(0.048)	(0.043)	(0.048)	(0.044)	(0.049)
sigma	2.717		2.733		2.727		2.721	
R^2	0.406		0.414		0.403		0.457	

Table 11: Tobit Index Values for Specifications 1 to 10

	Tobit Index 93	Tobit Index 95	Tobit Index 97	Tobit Index 00
Specification 1				
Constant	-14.990 *	-11.804 *	-15.909 *	-22.829 *
	(0.685)	(0.314)	(0.460)	(0.458)
Age	0.716 *	0.579 *	0.690 *	0.999 *
	(0.037)	(0.016)	(0.024)	(0.023)
Age squared	-0.007 *	-0.005 *	-0.006 *	-0.009 *
	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	-0.472 *	-0.306 *	-0.337 *	-0.375 *
	(0.056)	(0.026)	(0.033)	(0.037)
Sigma	0.065 *	0.048 *	0.046 *	0.049 *
	(0.004)	(0.002)	(0.002)	(0.002)
Log Likelihood	-16241.36	-67671.51	-58086.88	-54501.75
n	8487	31777	30956	29949
Specification 2				
Constant	-17.155 *	-13.568 *	-17.344 *	-25.104 *
	(0.698)	(0.314)	(0.460)	(0.453)
Age	0.716 *	0.579 *	0.691 *	1.008 *
	(0.038)	(0.016)	(0.024)	(0.023)
Age squared	-0.007 *	-0.005 *	-0.006 *	-0.009 *
	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	0.409 *	0.375 *	0.287 *	0.354 *
	(0.018)	(0.008)	(0.010)	(0.011)
Sigma	5.823 *	5.064 *	6.271 *	6.326 *
	(0.073)	(0.029)	(0.043)	(0.044)
Log Likelihood	-16352.61	-68045.15	-58284.30	-54712.22
Sepcification 3				
Constant	-15.250 *	-11.914 *	-15.953 *	-22.883 *
	(0.688)	(0.313)	(0.459)	(0.456)
Age	0.719 *	0.572 *	0.679 *	0.980 *
	(0.037)	(0.016)	(0.024)	(0.022)
Age squared	-0.007 *	-0.005 *	-0.006 *	-0.009 *
	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	-0.112	0.092 *	0.239 *	0.283 *
	(0.124)	(0.055)	(0.071)	(0.075)
(Education in years)^2	-0.006	-0.027 *	-0.066 *	-0.068 *
	(0.022)	(0.009)	(0.012)	(0.012)
(Education in years)^3	0.003 *	0.004 *	0.005 *	0.005 *
	(0.001)	(0.000)	(0.001)	(0.001)
sigma	5.681 *	4.974 *	6.195 *	6.234 *
	(0.071)	(0.029)	(0.042)	(0.044)
Log Likelihood	-16217.11	-67638.60	-58044.87	-54451.89

	Tobit Index 93	Tobit Index 95	Tobit Index 97	Tobit Index 00
Sepcification 4				
Constant	-12.357 *	-8.873 *	-9.953 *	-15.384 *
	(0.627)	(0.290)	(0.400)	(0.394)
Age	0.566 *	0.421 *	0.406 *	0.637 *
	(0.034)	(0.015)	(0.021)	(0.019)
Age squared	-0.006 *	-0.004 *	-0.004 *	-0.006 *
	(0.000)	(0.000)	(0.000)	(0.000)
Education in years	-0.126	0.101 **	0.120 **	0.148 **
	(0.113)	(0.051)	(0.062)	(0.066)
(Education in years)2	-0.018	-0.041 *	-0.054 *	-0.052 *
	(0.020)	(0.009)	(0.011)	(0.010)
(Education in years)3	0.004 *	0.004 *	0.004 *	0.004 *
	(0.001)	(0.000)	(0.001)	(0.000)
Union	5.809 *	4.518 *	6.900 *	6 807 *
	(0.169)	(0.069)	(0.090)	(0.091)
sigma	5.152 *	4.573 *	5.377 *	5.399 *
	(0.064)	(0.026)	(0.037)	(0.038)
Log Likelihood	-15630.18	-65577.12	-55142.01	-51741.25
Sepcification 5				
Constant	-11.864 *	-7.719 *	-9.294 *	-14.805 *
	(0.624)	(0.281)	(0.394)	(0.390)
Age	0.573 *	0.422 *	0.426 *	0.650 *
	(0.033)	(0.014)	(0.021)	(0.019)
Age Squared	-0.006 *	-0.004 *	-0.004 *	-0.006 *
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-1.162	-2.455 *	-1.863 *	-1.571 *
	(0.126)	(0.055)	(0.069)	(0.072)
Education in years	-0.162	0.027	0.087	0.128 **
	(0.112)	(0.050)	(0.061)	(0.065)
(Education in years)^2	-0.010	-0.023 *	-0.046 *	-0.048 *
	(0.020)	(0.008)	(0.011)	(0.010)
(Education in years)^3	0.004 *	0.003 *	0.004 *	0.004 *
	(0.001)	(0.000)	(0.001)	(0.000)
Union	5.618 *	4.145 *	6.545 *	6.490 *
	(0.168)	(0.067)	(0.089)	(0.091)
sigma	5.113 *	4.411 *	5.287 *	5.337 *
	(0.064)	(0.025)	(0.036)	(0.037)
Log Likelihood	-15587.65	-64596.38	-54778.97	-51504.35

	Tobit Index 93	Tobit Index 95	Tobit Index 97	Tobit Index 00
Sepcification 6				
Constant	-5.915 *	-4.008 *	-6.123 *	-10.044 *
	(0.620)	(0.281)	(0.409)	(0.400)
Age	0.530 *	0.425 *	0.478 *	0.644 *
	(0.031)	(0.014)	(0.020)	(0.019)
Age Squared	-0.005 *	-0.004 *	-0.005 *	-0.006 *
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-1.131 *	-2.314 *	-1.695 *	-1.463 *
	(0.119)	(0.053)	(0.066)	(0.069)
African	-5.017 *	-3.686 *	-4.325 *	-4.558 *
	(0.205)	(0.085)	(0.156)	(0.140)
Coloured	-2.350 *	-1.483 *	-0.902 *	-1.360 *
	(0.257)	(0.103)	(0.172)	(0.162)
Asian	-1.733 *	-0.609 *	-1.000 *	-1.045 *
	(0.329)	(0.144)	(0.264)	(0.244)
Education in years	-0.086	0.068	0.024	0.069
	(0.105)	(0.047)	(0.058)	(0.062)
$(\text{Education in years})^2$	-0.018	-0.038 *	-0.039 *	-0.041 *
	(0.019)	(0.008)	(0.010)	(0.010)
$(\text{Education in years})^3$	0.003 *	0.003 *	0.003 *	0.003 *
	(0.001)	(0.000)	(0.000)	(0.000)
Union	5.216 *	4.242 *	6.313 *	6.479 *
	(0.160)	(0.064)	(0.085)	(0.087)
sigma	4.786 *	4.176 *	5.039 *	5.072 *
	(0.060)	(0.024)	(0.034)	(0.035)
Log Likelihood	-15224.57	-63297.41	-53717.50	-50496.54

	Tobit Index 93	Tobit Index 95	Tobit Index 97	Tobit Index 00
Sepcification 7				
Constant	-4.424 *	-4.594 *	-5.561 *	-10.224 *
	(0.859)		(0.567)	(0.659)
Age	0.477 *	0.438 *	0.448 *	0.642 *
	(0.037)	(0.017)	(0.024)	(0.024)
Age Squared	-0.005 *	-0.004 *	-0.004 *	-0.006 *
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-1.137 *	-2.322 *	-1.691 *	-1.461 *
	(0.119)	(0.053)	(0.066)	(0.069)
African	-4.981 *	-3.671 *	-4.199 *	-4.530 *
	(0.207)	(0.086)	(0.157)	(0.141)
Coloured	-2.320 *	-1.465 *	-0.778 *	-1.333 *
	(0.259)	(0.103)	(0.173)	(0.163)
Asian	-1.756 *	-0.612 *	-0.888 *	-1.054 *
	(0.331)	(0.144)	(0.265)	(0.245)
Education in years	-0.104	1.228 *	1.133 *	0.812 *
	(0.351)	(0.166)	(0.218)	(0.238)
(Education in years)^2	-0.083	-0.279 *	-0.276 *	-0.195 *
	(0.064)	(0.028)	(0.038)	(0.037)
(Education in years)^3	0.008 *	0.016 *	0.015 *	0.011 *
	(0.003)	(0.001)	(0.002)	(0.002)
Age* Education	-0.001	-0.031 *	-0.031 *	-0.019 *
	(0.009)	(0.004)	(0.006)	(0.006)
Age* Education^2	0.002	0.006 *	0.007 *	0.004 *
	(0.002)	(0.001)	(0.001)	(0.001)
Age* Education^3	0.000 ***	0.000 *	0.000 *	0.000 *
	(0.000)	(0.000)	(0.000)	(0.000)
Union	5.196 *	4.224 *	6.273 *	6.474 *
	(0.160)	(0.064)	(0.085)	(0.087)
sigma	4.782 *	4.169 *	5.031 *	5.070 *
	(0.060)	(0.024)	(0.034)	(0.035)
Log Likelihood	-15217.66	-63246.37	-53684.10	-50483.62
Sepcification 8	-13.764 *	-9.296 *	-10.199 *	-16.315 *
Constant	(0.636)	(0.282)	(0.393)	(0.386)
	0.571 *	0.428 *	0.430 *	0.664 *
Age	(0.034)	(0.015)	(0.021)	(0.020)
	-0.006 *	-0.004 *	-0.004 *	-0.006 *
Age Squared	(0.000)	(0.000)	(0.000)	(0.000)
	-1.213 *	-2.483 *	-1.865 *	-1.547 *
Female	(0.130)	(0.057)	(0.070)	(0.073)
	0.335 *	0.315 *	0.126 *	0.191 *
Education in years	(0.017)	(0.007)	(0.009)	(0.010)
	5.617 *	4.178 *	6.684 *	6.656 *
Union	(0.173)	(0.069)	(0.090)	(0.092)
	5.271 *	4.504 *	5.334 *	5.397 *
sigma	(0.066)	(0.026)	(0.036)	(0.038)
LL	-15752.30	-65063.18	-54946.02	-51697.78

	Tobit Index 93	Tobit Index 95	Tobit Index 97	Tobit Index 00
Sepcification 9				
Constant	-6.043 *	-4.609 *	-6.350 *	-10.703 *
	(0.626)	(0.284)	(0.411)	(0.400)
Age	0.525 *	0.431 *	0.482 *	0.654 *
	(0.032)	(0.014)	(0.020)	(0.019)
Age Squared	-0.005 *	-0.004 *	-0.005 *	-0.006 *
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-1.148 *	-2.328 *	-1.690 *	-1.439 *
	(0.120)	(0.053)	(0.067)	(0.070)
African	-5.745 *	-4.098 *	-4.793 *	-4.937 *
	(0.198)	(0.085)	(0.155)	(0.139)
Coloured	-3.195 *	-2.000 *	-1.451 *	-1.786 *
	(0.249)	(0.102)	(0.170)	(0.161)
Asian	-2.109 *	-0.871 *	-1.340 *	-1.249 *
	(0.331)	(0.145)	(0.265)	(0.246)
Education in years	0.125 *	0.160 *	0.050 *	0.084 *
	(0.017)	(0.007)	(0.009)	(0.010)
Union	5.227 *	4.278 *	6.437 *	6.615 *
	(0.161)	(0.065)	(0.085)	(0.087)
sigma	4.841 *	4.234 *	5.074 *	5.110 *
	(0.060)	(0.024)	(0.034)	(0.036)
Log Likelihood	-15285.67	-63606.96	-53846.71	-50621.21

	Tobit Index 93	Tobit Index 95	Tobit Index 97	Tobit Index 00
Sepcification 10				
Constant	-4.828 *	-4.012 *	-5.441 *	-9.466 *
	(0.847)	(0.426)	(0.567)	(0.662)
Age	0.453 *	0.433 *	0.424 *	0.629 *
	(0.036)	(0.017)	(0.023)	(0.024)
Age Squared	-0.005 *	-0.004 *	-0.004 *	-0.006 *
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.952 *	-2.266 *	-1.631 *	-1.384 *
	(0.116)	(0.052)	(0.065)	(0.069)
African	-4.681 *	-3.396 *	-3.591 *	-4.093 *
	(0.208)	(0.087)	(0.158)	(0.143)
Coloured	-2.709 *	-1.425 *	-1.396 *	-1.912 *
	(0.285)	(0.108)	(0.184)	(0.176)
Asian	-0.889 *	-0.573 *	0.335	-0.667 *
	(0.348)	(0.147)	(0.269)	(0.252)
Education in years	-0.108	1.227 *	1.057 *	0.806 *
	(0.342)	(0.165)	(0.215)	(0.236)
(Education in years)^2	-0.082	-0.284 *	-0.267 *	-0.195 *
	(0.062)	(0.028)	(0.038)	(0.037)
(Education in years)^3	0.008 *	0.016 *	0.014 *	0.011 *
	(0.003)	(0.001)	(0.002)	(0.002)
Age* Education	0.001	-0.031 *	-0.030 *	-0.020 *
	(0.009)	(0.004)	(0.005)	(0.006)
Age* Education^2	0.002	0.007 *	0.007 *	0.004 *
	(0.002)	(0.001)	(0.001)	(0.001)
Age* Education^3	0.000 **	0.000 *	0.000 *	0.000 *
	(0.000)	(0.000)	(0.000)	(0.000)
Union	4.794 *	4.147 *	6.110 *	6.469 *
	(0.155)	(0.063)	(0.083)	(0.087)
Western Cape	1.866 *	0.025	1.676	0.718 *
	(0.259)	(0.109)	(0.147)	(0.151)
Eastern Cape	-0.922 *	-1.691 *	-1.812 *	-2.036 *
	(0.235)	(0.095)	(0.135)	(0.134)
Northern Cape	0.079	-0.790 *	0.922 *	-0.151
	(0.503)	(0.147)	(0.177)	(0.180)
Free State	-3.100 *	-0.114	1.038	-0.532 *
	(0.417)	(0.103)	(0.129)	(0.138)
KwaZulu-Natal	-0.094	-0.523 *	-1.194 *	-0.826 *
	(0.208)	(0.090)	(0.118)	(0.118)
Northwest Province	2.034 *	-0.341 *	0.117	-0.774 *
	(0.223)	(0.110)	(0.127)	(0.128)
Mpumulanga	2.270 *	-0.536 *	0.752 *	-1.090 *
	(0.224)	(0.106)	(0.132)	(0.143)
Northern Province	0.191	-1.420 *	-0.112	-1.728 *
	(0.237)	(0.118)	(0.137)	(0.140)
Sigma	4.625 *	4.128 *	4.942 *	5.017 *
	(0.058)	(0.024)	(0.033)	(0.035)
Log Likelihood	-14985.02	-62975.37	-53247.25	-50255.31