Using Panel Data to Estimate the Returns to Schooling in South Africa

Caitlin Miles (MLSCAI001)

Supervisor: Reza Daniels

School of Economics, University of Cape Town

A thesis submitted in partial fulfillment of the requirements for a degree of Masters of Commerce in Applied Economics

Word Count: 19035

Returns to schooling have typically been estimated with cross-sectional data. However, these studies are fraught with difficulties arising from the endogeneity of education. Individual effects that cannot be measured, such as ability and family background, cause bias in the estimates because they are correlated with education. A panel data approach is thus potentially superior to a cross-sectional one, in that it allows the individual effects to be eliminated with time-differencing. However, time-invariant regressors, such as education, cannot be identified under these time-differencing techniques. This paper therefore uses a Generalized Instrumental Variables method that was developed by Hausman and Taylor (1981) to estimate returns to schooling under a panel data context. This approach both controls for endogeneity bias and allows the identification of time-constant regressors, in this case, education. The returns to schooling under this estimation method are approximately 21% for South African individuals who are consistently employed from 2008-2013.

Contents

List of Figures

List of Tables

Plagiarism Declaration

1. I know that plagiarism is a serious form of academic dishonesty.

2. I have read the document about avoiding plagiarism; I am familiar with its contents and have avoided all forms of plagiarism mentioned there.

3. Where I have used the words of others, I have indicated this by the use of quotation marks.

4. I have referenced all quotations and other ideas borrowed from others.

5. I have not and shall not allow others to plagiarise my work.

Date: 18 May 2015

Signature: 

Signature removed

1 Introduction

Returns to schooling refer to the private wage effect from investing in additional education. Estimating and analysing these returns is "probably the most explored and prolific area in labour economics" (Battistin, De Nadai and Sianesi, 2012:2). This area is of interest due to its policy relevance. Quantifying the returns to education is important for policy debates on the potential of different levels of education to reduce poverty. In particular, with increased enrollment in schooling across developing countries, it is of interest for researchers to compare the relative costs and benefits of this schooling.

Additionally, the interest in this topic arises from the unresolved debate as to the size of these returns. Econometrically, there is great difficulty in extrapolating from the positive correlation between wages and education to a causal inference about education's effect on wages. The bulk of the debate has thus concerned disentangling this relationship from confounding factors and isolating the effect of schooling on earnings alone. As Heckman, Lochner and Todd (2006:1) gloomily put it, these returns are "widely sought after and rarely obtained".

Returns to schooling have typically been estimated with cross-sectional data. However, these studies are fraught with difficulties arising from the endogeneity of education. Individual-specific effects that cannot be measured, such as ability and family background, cause bias in the estimates because they are correlated with education. A panel data approach potentially presents an advantage to a cross-sectional one, because observed variables are captured at several points for each individual. Using this data could thus allow the confounding individual effects to be eliminated, assuming they are time-constant.

However, in the returns to schooling literature, little analysis has been conducted with panel data. Moreover, to this author's knowledge, there are no studies that look at returns to education with panel data in South Africa. This has likely been due to the lack of panel data in the past. However, the recent release of the National Income Dynamics Study (NIDS) represents the first national panel survey in South Africa. With three waves so far, there is thus scope to use this panel dataset to estimate returns to schooling. This paper will act as the first attempt at exploring this topic with a panel dataset in South Africa.

There are numerous complications to a panel data analysis of returns to schooling. One of the main concerns is that education of the current workforce does not have sufficient variation to identify a valid coefficient. This is due to the fact that the education of most working individuals is time-constant over short panel studies. A further difficulty is the endogeneity concerns of education, which imply that an instrumental technique is required to isolate

exogenous variation in education. However, as with any instrument, there are difficulties in proving that the instrument is orthogonal to the error term. This paper will detail this sequence of difficulties, and will thus be frank in its assessment of the advantages of using panel data in estimating returns to schooling.

The main estimation method that this paper will employ is Hausman and Taylor's (1981) Efficient Generalized Instrumental Variables approach. These authors developed this method with an application to returns to schooling. Its foremost benefit is the use of instruments internal to the dataset. This removes the need for hard-to-come-by natural experiments or other external instruments, such as a change in education policy. If the internal instruments are chosen correctly, they can act as sufficient exogenous variation in education to identify the returns to schooling. This approach therefore both controls for endogeneity bias and allows the identification of time-constant regressors, in this case, education.

The first section of this paper will set out the background theory to returns to schooling. It will discuss the human capital perspective that underlies the estimation, and will detail recent thinking on this theory. Issues around the empirical estimation of the model will then be considered, from specification concerns to the sources of bias in the model. The various methods that have been used to deal with endogeneity bias will be discussed, and the use of panel data in this light will be judged in greater depth. This section will also summarize the current education system in South Africa, and it will highlight the estimation methods that have been employed to estimate returns up to now.

The second section of this paper will act as the methodology section. It will consider the data in more depth, and will present a descriptive analysis of the sample of working individuals. Consideration of missing income data for the sample will also be given. Finally, it will set out in detail the Hausman-Taylor (HT) estimation method that will be taken. The third section will present the results of this analysis. It will start with the initial cross-sectional analysis of returns to schooling. It will then progress to the panel data analysis of returns to schooling, and will implement the Hausman-Taylor estimation method.

The final section will discuss the results in more depth. It will consider the limitations of the analysis both in terms of covariates and sample size. It will also highlight the difficulty in choosing which internal instruments are valid, and the resulting sensitivity of the results to this choice. As such, this paper will present a best attempt at using panel data to estimate returns to schooling, but will acknowledge both the sensitivity and the sample-specific nature of these results.

2 Background

2.1 Theory on Returns to Schooling

The return to schooling is defined as the discount rate that equates the net benefits of investing in education with the net costs incurred for this schooling. These rates have been extensively estimated since Gary Becker and Jacob Mincer established the groundwork for Human Capital Theory in the 1960s and 1970s. Prior to this, little analysis into the economics of education had been conducted. The key insight of Human Capital Theory is that education represents the stock of productive skills that an individual acquires. Increased productivity in turn augments lifetime earnings. As such, education can be seen as a driving force behind the distribution of incomes within a population. Education is thus termed "human capital" under this theory to represent the fact that it is a investment yielding future economic gains.

The starting point of this theory was the application of neoclassical price theory to the concept of education. Becker (1964) set out the seminal model where the decision to invest in education is modelled as any other factor of production. Eduction increases future and lifetime earnings. But its costs are borne in the present, both in the form of direct fee costs and indirect foregone earnings. As such, a rational individual will choose the education level that equates the marginal benefit of the s^{th} year of education with its marginal costs. This equilibrium can be set out as follows,

$$\sum_{t=1}^{T-S} \frac{w_s - w_{s-1}}{(1+r_s)^t} = w_{s-1} + c_s \qquad (1)$$

where s refers to years of schooling, T the number of years until retirement, w the wages, and c represents the cost of education. The left-hand side of (1) represents the present discounted value of wages in year s. The right-hand side denotes the foregone wages and the cost of going to school in year s. If the individual's rate of return r_s is greater than the market rate of return, then the individual will choose to invest in more schooling (assuming there are no constraints to borrowing). As such, an equilibrium is formed when the market rate is equivalent to the individual's return, which is then used as the rate to discount the future earnings stream (Harmon, Oosterbeek and Walker, 2000).

Mincer's (1974) paper further entrenched a human capital approach. He followed on from the Becker equilibrium to derive an equation for modelling the returns to schooling. If one

assumes that T is large, then the left-hand side of the equation can be approximated by $\frac{w_s - w_{s-1}}{r_s}$. If c_s is then assumed to be small enough, the following approximation can be made:

$$r_s \approx \frac{w_s - w_{s-1}}{w_{s-1}} \approx log(w_s) - log(w_{s-1}) \qquad (2)$$

This implies that the return to an additional year of education is approximated by the difference in wages between leaving school at year s or at year $s-1$. Mincer (1974) then makes several other assumptions, such as the proportion of the time period t spent acquiring human capital skills declines linearly with experience once the individual has completed school. This allows the derivation of his classic semi-logarithmic earnings equation,

$$logEarnings = \beta_o + \beta_1 School + \beta_2 Experience + \beta_3 Experience^2 + \epsilon \qquad (3)$$

where the right-hand side variables are years of education and experience. According to Mincer (1974), earnings are thus explained by the amount of human capital, which is largely captured by education and experience. Both schooling and on-the-job training are thus key components of an individual's human capital (Psacharopoulos and Patrinos, 2004). Experience is included as a quadratic to reflect the concavity of the earnings function - earnings rise with experience but at a decreasing rate.

The important implication of (3) is that B_1 represents the return to schooling. As is the case for any independent variable in a multiple regression, the beta coefficient on schooling can be interpreted as the change in log earnings from an additional year of education, holding all else constant. This accords with (2), where the return to education is approximated by the difference in log earnings from the s^{th} year of school. Putting these two propositions together, it thereby implies that the beta coefficient on schooling approximates the rate of return to education.

However, as Gustafsson and Mabogoane (2012) point out, the typical Mincerian approach to estimating these returns fails to account for the costs of educational investment. The approximation in (2) was only possible under the assumption that c_s is sufficiently small. As such, these beta coefficients are best termed "wage effects". Alternatively, they can be interpreted as rates of return if school is free and children are unable to work during their school-age (Card, 1999). Estimating returns to schooling with a Mincerian approach has become the convention in the literature, despite this underlying assumption.

In recent years, there have been several developments in the field of human capital theory. Much of this work has consisted of determining the mechanism that makes schooling productive and thereby increases earnings. Rosenzweig (2010) argues that schooling augments productivity through two avenues: it imparts specific knowledge, and it improves an individual's skills in acquiring knowledge. The latter feature has been termed "learning effects" and captures the hypothesis that education improves an individual's access to information and his ability to understand and process new information (Jones, 2001 and Rosenzweig, 1995).

Learning effects have been subject to much study, particularly based on the Green Revolution in India. This research finds that people with greater education adopt new technology crops at a faster rate. These individuals tend to be more forward-looking and thereby demonstrate greater learning effects (see Besley and Case, 1993 and Bandiera and Rasul, 2006). Outside of farming, Dupas (2014) finds that educated individuals are more likely to adopt anti-malarial bed nets over the long-term. As such, education seems to be complementary to new technology adoption, and it is this characteristic that increases the individual's productivity.

Other work in the field of human capital theory has concerned the actual outcome variable for returns to schooling. Typically, an individual's take home pay from a formal job is used as the dependent variable. However, Oreopoulos and Salvanes (2011) have argued that the typical earnings variable does not account for indirect benefits such as medical insurance, pension contributions or stock options. They argue that returns to schooling are 10-40% higher when including these benefits. Furthermore, some scholars have argued that in low-income countries, wages are not an appropriate measure of the returns to schooling. Such an approach neglects self-employment and the informal sector that dominates the economies of these countries. Often in these contexts, wages are paid in kind, such as the provision housing or food (Rosenzweig, 2010). As such, using permanent wages as a measure of the outcome variable may not be completely adequate.

Furthermore, there is increasing research into the social returns of education. This includes the non-pecuniary benefits of schooling, such as its effects on long-term job satisfaction or on mental health (Oreopoulos and Salvanes, 2011). Additionally, this research has expanded into attempts to quantify the effects of schooling on wider society, through mechanisms such as crime reduction, fertility decreases, child wellbeing and citizen participation (Hanushek and Woessmann, 2008). As such, this field has seen an increasing shift in focus to education's effect on the individual and societal welfare overall.

In summary, estimating returns to schooling is based on an underlying human capital perspective. According to this theory, education is acquired to increase the individual's productivity

and subsequent earnings. This productivity increase is largely driven by greater technology adoption and learning effects. The individual chooses the level of education that equates the present value of his future earnings with the direct and indirect costs of this schooling. The return to education is then the underlying discount rate that is required for this equilibrium to hold. Under the assumptions of zero-cost schooling and that wages adequately capture the benefits accrued to the worker, this return can be estimated as the coefficient on schooling when earnings is the dependent variable.

2.2 Empirical Estimation of Returns to Schooling

There are numerous complexities to consider when using a Mincer equation to estimate returns empirically. These complexities can be divided up into two categories. The first concerns specification issues - factors around the measurement of the variables in the Mincer equation and their functional form. The second issue is the sources of bias in the equation. If any regressors are correlated with the error term, the coefficients will be biased and a valid interpretation of them cannot be made. Each issue will now be discussed in more depth, and potential strategies to deal with them will be considered.

2.2.1 Specification Issues

Mincer's canonical function makes some fairly strong estimation assumptions. Schooling is measured as years of education. The Mincer equation thus assumes linearity – each additional year of education has an equivalent proportional effect on wages. Hungerford and Solon (1987) were among the first to argue for the existence of nonlinearities, whereby certain years cause a jump in wages (such as the completion of primary or high school). This has been termed the "sheepskin effect" in the literature. However, Card (1999) argues that even though this assumption of linearity seems unrealistic, a linear function generally fits surprisingly well. This is confirmed by Harmon et al. (2000), who find that a linearity assumption is hard to reject in the United Kingdom. They explain that a three-year degree tends to generate three times the returns to an A-level qualification, thereby according with an overall linearity assumption.

Additionally, the Mincerian equation assumes that years of schooling is a true measure of education. This is unlikely to be valid if there are high levels of grade repetition in the sample, or if the quality of education is low (Card, 1999). As such, particularly in developing

countries, years of education may not be an adequate measure of the true underlying level of human capital. Dummy variables for the completion of a primary, secondary, or tertiary qualification can be used as a secondary measure of education. This approach avoids the linearity and grade repetition criticisms leveled against the continuous measure. However, if the survey question has been phrased to elicit the years of education, converting these responses to dummies may yield other measurement errors.

A further issue is the measure of experience in the Mincer equation. Few datasets have this information and it is tricky to get an accurate measure of it. As such, many scholars use a proxy of $experience = age - years\,of\,schooling - 6$. However, if there is high proportion of grade repetitions or high levels of unemployment, then this proxy will tend to overestimate experience, and it will be subject to serious measurement error (Garcia and Montuenga, 2005). Age is sometimes used as an alternative to experience because it is likely to have less errors. However, as Harmon et al. (2000) point out, this will mean that the coefficient needs a slight adjustment to be comparable to Mincer equations where experience is used. This is because there is a difference in what is being held constant in a ceteris paribus interpretation. In reality, the adjustment that is required tends to be fairly small (-0.0005 on average for men), and so it is often ignored (Harmon et al., 2000).

2.2.2 Sources of Bias

A further significant concern in the empirical estimation of the Mincer equation is that there are several biases in an OLS estimation approach. Firstly, selection bias may cause inflated returns to schooling, as the working sample is likely to be a non-random subset of the labour force. Those selected into the workforce may be systematically different to the unemployed and thus obtain higher earnings than the unemployed. As such, investment into education may not, in reality, yield as attractive an investment as implied by the OLS estimates.

Falaris (1995) suggests that this is particularly a problem in developing countries where selectivity tends to be ignored in research, but where high unemployment rates are common. In Falaris's study of Venezuelan women, he finds that the rate of return when adjusting for selectivity is 8.6%, down from the OLS estimate of 12.1%. However, it must be noted that if the question of interest concerns the returns to schooling *of those that are employed*, then selection bias can be ignored. This is the assumption made by papers that neglect to control for employment determination.

A second significant source of bias in OLS estimation is that of omitted variables. The choice of educational attainment is correlated with individual unobserved characteristics.

This implies that education is not merely a reflection of the productivity of an individual, but it additionally reflects these underlying characteristics. As such, a coefficient on education would capture both the effect of education on wages, and these unobserved variables. Unfortunately the "gold standard" of Randomized Control Trials cannot be drawn upon in this case because schooling levels cannot be randomly assigned across the population. These individual unmeasured characteristics continue to determine schooling choices and thereby complicate the analysis (Card, 1993).

In particular, much attention has been given to unobserved ability that influences an individual's earning power in the labour market. Ability captures characteristics such as intelligence, work ethic and levels of motivation (Griliches, 1977). Bowles, Gintis and Osborne (2001) have argued that it even incorporates the communication skills and general attitude of an individual. It may be that individuals with higher earnings capacities due to these ability characteristics are merely acquiring more schooling, rather than higher earnings being caused by greater education (Card, 1999). In a simple Mincer equation, there will thus be bias if ability has both a direct effect on wages and is correlated with education. This bias can be represented as follows (Wooldridge, 2012: 88):

$$plim\hat{\beta}_{OLS} = \beta_{true} + \frac{Cov(Schooling, Ability)}{Var(Schooling)} \qquad (4)$$

The direction of this bias will depend on the nature of the interaction between education and ability (Arias et al., 2002). If they are complements, then more able individuals will acquire more education, perhaps because of lower marginal costs to this acquisition. The interaction between ability and education would thus be positive, which leads to an overall upward bias and an overestimate of the returns to schooling (Card, 2001). If they are substitutes, then this relationship is negative. More able individuals will choose less education, perhaps because they can meet their desired wage level with relatively less education. Scholars have, however, typically assumed an upward bias from ability, thereby making the complement assumption.

Family background is the second omitted variable. It captures the family learning environment, or the family's extent of ambitiousness and work ethic. Family background is likely to play an important role in determining communication abilities, resourcefulness and general attitude to work. This omitted variable will also capture the family's labour market connections that influence work attainment or earnings. These features of family background are likely to interact with education, and will therefore bias the estimated coefficient on schooling (Heckman and Hotz, 1986). Behrman and Wolfe (1984) used a differencing approach for

sister pairs in Nicaragua. They found that controlling for family background in this manner caused a drop in the OLS estimates of returns to schooling by one-fourth. Omission of family background effects can therefore severally bias estimates.

In summary, there are several complexities that plague the estimation of returns to schooling. Firstly, there are issues surrounding the measurement of education. Years of schooling may not always capture nonlinear jumps in returns to schooling, and may not be the best representation of human capital accumulation. Experience is also often subject to poor measurement. Second, and more importantly, is the issue of bias. Selection bias may be a problem if returns to schooling are intended to be nationally representative but fail to take into account the selection into the labour force. An even more severe concern is that of omitted variable bias. Individual unobserved characteristics that encompass family background and ability are correlated with the choice of educational attainment. This implies that the Mincer coefficient is not solely capturing the productivity effect of education on wages. Consideration of how to account for this endogeneity problem will now be made.

2.2.3 Approaches to Dealing with Omitted Variable Bias

Finding a solution to omitted variable bias can be tricky. Some researchers have made use of IQ scores as a proxy for ability. However, IQ scores may not adequately capture the characteristics, such as ambitiousness or resourcefulness, that influence earnings. As Griliches (1977) notes, ability has little to do with IQ, and is better described as individual motivation or even energy levels. Furthermore, IQ scores may not fully capture the influence of family connections or family work ethic that nevertheless play an important role in earnings determination.

An approach to better net out family effects could be an estimation of within-family differences. If individual i belongs to family j, the family-average could be subtracted from each variable: $Earnings_{ij} - \bar{Earnings}_i = (X_{ij} - \bar{X}_i)\beta + \varepsilon_{ij} - \bar{\varepsilon}_j$. This would eliminate any unobserved effects that are common to the family. However, such an approach may conversely fail to capture individual ability factors that do not run in the family (Arias et al., 2002). Using identical twin data to do a differencing approach would appear to be the most useful in netting out both ability and family background characteristics. Identical twins share genetic material and a family background, and are thereby likely to have a common unobserved effect. However, the availability of identical twin data is typically limited. Furthermore, twins may be systematically different to the wider population, and thus this approach fails to be highly useful or generalisable (Harmon et al., 2000)

More recently, scholars have looked to a variety of instrumental variables (IVs) to identify returns to schooling. The set-up for an instrumental approach is as follows, where Z denotes the matrix of instruments (Wooldridge, 2002):

$$logEarnings_i = X_i\beta + \gamma School_i + \epsilon_i \qquad (5)$$

$$School = Z_i'\alpha + v_i \qquad (6)$$

Equation (6) is used to compute a predicted value for schooling, which then replaces schooling in the structural earnings equation (5). This is known as a Two-Stage Least Squares instrumental approach. In order for this method to be valid, two conditions are required (Harmon et al., 2000). Firstly, the instrumental variables in Z must be correlated with schooling to ensure that the effect of schooling on wages can be captured. This is known as the instrumental relevance requirement. Secondly, the schooling equation cannot be correlated with the unobserved variables in (5). It needs to capture the variation in schooling that is exogenous to the error term in the earnings function. As such, instrumental variables in Z cannot be concurrently contained within X; they cannot have an effect on earnings outside of the schooling equation. This is known as the instrument exogeneity or orthogonality condition.

Researchers have tended to use supply-side variables as a way to meet the instrumental conditions in the analysis of returns to schooling. For example, Angrist and Krueger (1991) used the individual's quarter of birth as an instrument for schooling. This is correlated with years of schooling because the quarter of birth allows individuals born earlier in the year to more quickly reach the minimum school-leaving age, and thereby obtain slightly less schooling. But it is unlikely that the quarter of birth is correlated with ability and it should thereby meet the instrument exogeneity requirement. Another common supply-side instrument that is employed is the geographical proximity to schools (Card, 2001). Again this variable is likely to be correlated with years of education but uncorrelated with ability.

However, the instrument exogeneity requirement can only be motivated theoretically - it is impossible to test empirically if the instrument is uncorrelated with the error term. As such, even the most innovative of instruments is subject to criticism. For example, Bound and Jaeger (1996) argue that there are sociological reasons why quarter of birth is likely to be correlated with family background. As such, this landmark instrumental variable conceived of by Angrist and Krueger may not be as promising as initially thought. Proving that the instrument meets the exogeneity condition of being unrelated to ability or family background

is thus an ongoing problem in returns to schooling studies (Card, 2001).

A further concern with using an instrumental approach is that the resulting estimates tend to be 20-40% higher than the OLS estimates (Card, 1999). This is contrary to expectation: an instrumental approach should deal with the upward bias caused by the omitted variables, and the estimated return to schooling should thus be lower than the OLS equivalent. There are several possible explanations to this puzzle. The first is that the omitted variable bias actually acts in a downward manner: more able individuals choose less education because they are able to reach a target wage without it. This could operate through the marginal cost of schooling: individuals with higher ability have a greater marginal cost in attaining education, and therefore reduce their schooling attainment. However, for most scholars, the plausibility of this theory is questioned. It seems more reasonable to assume that able individuals choose higher education (Card, 2001).

A second explanation is that measurement error in the education variable causes a downward bias, which outweighs the omitted variable bias (Griliches, 1977). The measurement error is assumed to be classical in nature, and therefore leads to an attenuation bias. As such, an instrumental approach would correct for this measurement error, and thereby lead to increased estimates of returns to schooling. However, with a categorical variable, such as education, there is now increasing consensus about the nonclassical nature of this measurement error. Individuals with a low education level cannot underreport by much, and individuals with high education levels cannot overreport (Kane, Rouse and Staiger, 1999). As such, the measurement error is mean-regressive, and the bias is not necessarily an attenuation bias. Furthermore, scholars have tended to find that when correcting for known measurement error, the estimates fall, thereby suggesting that measurement error in the education variable leads to upward bias (Hertz, 2003). As such, this explanation for the IV estimates has been largely disregarded.

The most plausible cause of the high IV estimates is set out by Card (1999) and Harmon et al. (2000). These scholars start with the assumption that individuals with the lowest schooling tend to have the highest marginal returns to schooling. Furthermore, it is typically these individuals with the lowest schooling that are targeted by the interventions that are used in IV approaches. For example, a government intervention that builds schools will ensure that new schools are placed in isolated areas, close to individuals who would otherwise obtain low schooling. As such, the intervention has the greatest effect on those with the highest marginal return to schooling. This leads to an upward bias in the IV estimates, because the subgroup that dominates are those that have above-average marginal returns to schooling.

In summary, the typical solutions to the endogeneity problems of education are still subject to problems. A proxy for ability or a within-family differencing technique fail to fully control for the omitted variables. Instrumental variables that make use of a supply-side feature of schooling may still be correlated with the error, and therefore be rendered invalid. Furthermore, instrumental results of returns to schooling tend to be even higher than the OLS results, suggesting that the instrument may capture the returns to schooling for particular subgroups that have above-average returns. A final approach to controlling for endogeneity, in the form of panel data, will now be discussed in some depth.

2.2.4 Panel Data as a Solution to Omitted Variable Bias

When using a longitudinal approach, the omitted variables are modelled as the individual fixed effect. This implies an assumption that the unobserved effects are time-invariant and individual-specific. In other words, each individual has a specific unobserved ability or motivation level that is constant over time. This model can be represented as follows,

$$Earnings_{it} = X_{it}\beta + c_i + u_{it} \qquad t = 1, ..., T; \;\; i = 1, ..., N \qquad (7)$$

where $Earnings_{it}$ is the wage of individual i in time t, X_{it} is the vector of observed individual characteristics including schooling; c_i is the individual time-invariant unobserved heterogeneity; and u_{it} is the individual time-varying unobserved disturbance. The further important assumption is that c_i is correlated with the schooling regressor: $E[c_i \mid Schooling_{i1}, Schooling_{i2}...Schooling_{iT}] \neq 0$. This captures the correlation of the omitted variables, family background and ability, with education.

There are two main estimation approaches when using panel data. The first is random effects. This approach assumes that there is time-constant individual heterogeneity that is in operation, but that this heterogeneity is not correlated with the regressors. In other words, any underlying ability or family background characteristics do not determine educational attainment: $E[c_i \mid Schooling_{i1}, Schooling_{i2}...Schooling_{iT}] = 0$. This approach therefore does not attempt to eliminate the individual effect, because this effect is believed to not induce bias.

However, random effects does adjust for the covariance structure when an individual effect is present. If the composite error term is $v_{it} = c_i + u_{it}$, and the variance-covariance matrix is

$E[v_i v_i] = \Omega$, then the presence of c_i causes the error matrix to be non-diagonal. The time-constancy of c_i implies an obvious serial correlation of the error terms over time. Additionally, it can be shown that c_i will also lead to heteroskedasticity (Wooldridge, 2002). As such, any estimated standard errors will not be correct due to this underlying heteroskedasticity and serial correlation. Random effects is a form of feasible generalised least squares, because it adjusts for these incorrect error terms. It first derives an estimate of the error matrix, $\hat{\Omega}$, which is then used to weight the model:

$$\beta_{RandomEffects} = \sum_{i=1}^{n} \left(X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Omega}^{-1} Y_i \right) \qquad (8)$$

The problem with a random effects approach is that the estimates will be biased if the unobserved effect is indeed correlated with the regressor. As discussed previously, it is likely that education is correlated with individual ability and family background, and random effects estimates will thus be rendered invalid. This leads on to the second main estimation approach with panel data: that of time-differencing. This approach allows the elimination of the time-constant unobserved effect by differencing the variables over time. With first differencing, the lag of the variable is subtracted, thereby ensuring that the time-constant individual effect is eliminated:

$$Earnings_{it} - Earnings_{i,t-1} = (X_{it} - X_{i,t-1})\beta + u_{it} - u_{i,t-1} \dashrightarrow \triangle Earnings_i = \triangle X_i \beta + \triangle u_i \qquad (9)$$

The new error term is thus mean independent of the differenced schooling regressor and the problem of endogeneity has therefore been dealt with. Similarly, with fixed effects estimation, the individual's specific average of the variable over time is subtracted. Again, this causes the time-constant unobserved effect to be eliminated, and thus controls for the endogeneity of schooling:

$$Earnings_{it} - \bar{Earnings_i} = (X_{it} - \bar{X}_i)\beta + u_{it} - \bar{u}_i \dashrightarrow \ddot{Earnings_i} = \ddot{X}_i \beta + \ddot{u}_i \qquad (10)$$

However, there is a fundamental problem with a time-differencing approach applied to returns to schooling. All regressors are assumed to be correlated with the individual effect and are thus differenced. The unavoidable problem is that time-constant regressors are concurrently eliminated. In particular, education is constant for individuals who have completed schooling. A time-differencing approach will thus eliminate this covariate, and its coefficient cannot be identified. And, since education is the variable of interest for a returns to schooling analysis, a time-differencing approach is therefore rendered futile.

If, however, the panel does include some workers who have changed their education over the waves, then a differencing approach may allow for estimation. However, workers who return to school are likely to be systematically different from those who do not, resulting in a biased estimate of returns for the labour force as a whole (Harmon and Walker, 1995). Additionally, measurement error tends to be exaggerated in a panel data context. This is because a small portion of the sample who changed their schooling are used to identify the effect, and any error in this reported schooling will thus tend to be exaggerated (Swaffield, 2001). As such, attempting to use a sub-sample of those who changed their education is not an adequate solution.

Hausman and Taylor (1981) developed an approach to dealing with the problems of using panel data to estimate returns to schooling. Their method can be seen as an intermediate approach between random effects and fixed effects. Whilst random effects treats all variables as exogenous, and fixed effects effects assumes all variables are endogenous, the HT approach chooses some variables to be exogenous and others as endogenous. The authors then exploit internal instruments in a way that allows them to both deal with the endogeneity of education and identify the time-invariant regressor. They divide the regressors into either a time-constant or a time-variant category. They then use the exogenous, time-varying regressors as instruments for the endogenous, time-constant variables. The authors also include a weighting procedure into their approach to ensure efficient estimates. The overall HT method is thus known as Efficient Generalized Instrumental Variables, and it yields consistent and asymptotically efficient estimators (Wooldridge, 2002).

In their original 1981 paper, Hausman and Taylor use a sample of workers for whom education is time-constant by limiting the sample to those not enrolled in education. Age, bad health and unemployment in the previous year act as the time-varying exogenous variables. The time-constant exogenous variables are race and union status, whilst the time-invariant endogenous variable is education. They use only two years from the Panel Study on Income Dynamics to minimize problems of serial correlation. With this approach, they find that the estimates of the returns to schooling increase from 7% with OLS, to 12-13%, suggesting that the individual heterogeneity bias was in a downward direction.

Several other studies have used the HT approach to estimate returns to schooling. Arcand, d'Hombres and Gyselinck (2004) similarly find that the HT technique yields estimates of returns to education that are higher than the OLS counterparts for two rounds of the Vietnam Living Standards Survey. Wright (1999) finds that OLS under-estimates returns by a significant margin when he uses the British Household Panel Survey. Garcia and Montuenga (2005) considered the returns to wage earners versus self-employed workers. They use a panel

dataset for Portugal and Spain, and find that secondary education yields the highest return for the self-employed, whilst wage earners benefit most from a tertiary qualification.

To sum up, panel data appears to be a useful approach to dealing with omitted variable bias. Assuming that the unobserved effect is time-constant, it can be eliminated with some form of differencing over time. However, this approach simultaneously eliminates the variable of interest, namely education. The HT method presents a way to recover this coefficient through an instrumental procedure. It also corrects for errors correlated over time by a weighting method. As such, it can yield consistent and efficient estimates of returns to schooling.

2.3 Education in South Africa

The South African education system is characterised by low "effective enrollment" (Spaull and Taylor, 2013). There is high enrollment in school, but the quality of learning is low. Using the Demographic and Health Surveys, together with the Southern and Eastern African Consortium for Monitoring Educational Quality III, Spaull and Taylor (2013) find that 98% of grade six aged children are enrolled in school. However, only 71% are functionally literate, and only 59% are functionally numerate. The poor standard of education was confirmed in the World Economic Forum's 2014 Information Technology report. Using perceptions from business leaders, the quality of South Africa's education system was ranked 146th out of 148 countries. Furthermore, it came last in the mathematics and science rankings. However, enrollment in secondary schooling is high, and South Africa ranks 28th out of 148 (Bilbao-Osorio, Dutta and Lanvin, 2014).

Although there is high enrollment in school, there is a high drop-out rate before secondary schooling is complete. Spaull (2013) finds that almost 50% of students drop out over the course of school. Furthermore, students also tend to drop out of the more difficult subjects, and there has been a 56% fall in the number taking pure mathematics between 2008 and 2011. Similarly, Branson, Kekana and Lam (2013), when using NIDS data, find that education is almost universal up to the age of 15. However, failure to complete the Matric certificate is a severe problem, and 35% of grade 9 learners in 2009 were no longer enrolled in school in the third wave of NIDS in 2012 (Branson et al., 2013).

The dismal facts about the quality of South African schooling are particularly pertinent given the amount of government spending on education. The government allocates one-fifth of its budget to education, higher than most other developing countries (Spaull, 2013). However, recent research has revealed a trend in enrollment into low-fee private schools. The Centre

for Development and Enterprise released a report in 2013 ("Affordable Private Schools in South Africa") detailing the growth in the number of independent schools between 2000 and 2010 by 44%. Conversely, public schools have declined over the same period by 9%.

The context for schooling in South Africa is therefore one of poor quality education in government schools despite the large amount that is spent on education. This yields high drop-out rates as pupils are not able to reach the standard required to pass the final-year exam. Against this background, looking at returns to schooling is therefore of interest in South Africa. Returns to schooling studies can yield policy suggestions about whether there is ongoing gains to spending on schooling, and they can shed light on the focus that government should give to particularly levels of educational attainment.

2.3.1 Returns to Schooling in South Africa

The returns to schooling in South Africa have been widely researched. However, the majority of this research has applied OLS to data of a cross-sectional nature, and little attempt has been made to correct for the sources of OLS bias. Much of the literature in this area has focused on two factors: firstly, the differentials in the returns to education by race, and secondly, the shape of the returns to education schedule.

In terms of the first focus, the literature has typically found that the education return is significantly higher for white than black workers. Bhorat (2000) estimates OLS equations by race using the 1995 October Household Survey. Even when controlling for occupational skills, he finds higher returns to whites for primary, secondary and tertiary education. For example, whites obtain a 26% return on a year of tertiary education, whilst this return is only 16% for blacks. This result appears to be due to the historical factors that influence the quality of the schooling institutions and universities for the different races (Bhorat, 2000).

Keswell (2004) uses decomposition methods to tease out racial wage differences using the Project for Statistics on Living Standards and Development (PSLSD) and the 2001/2002 Labour Force Surveys surveys. He finds that it is no longer a strong direct racial discrimination effect that causes earnings deviations, but rather the valuation of the labour market of education differs between the races. He estimates that these differences in returns to education between whites and blacks account for 40% of the overall wage differential. Similar to Bhorat (2000), he argues that this is likely due to the continuation of unequal quality of education and job segregation.

Regarding the second focus of the South African literature, the return to education relationship appears to be highly convex for the South African case. This implies that the rate of return to an additional year of schooling increases with the level of education completed. Bhorat (2000) finds a tertiary rate of return of 16%, but a primary rate of return of only 4% for African workers. Similarly, Fryer and Vencatachellum (2005) find no returns to primary education when looking at a sample of Black women. A wage premium to education only appears after two years of secondary schooling. The convex shape in the returns schedule is confirmed by Keswell and Poswell (2004), who introduce a cubic polynomial in education to the standard Mincer equation. The significance of this polynomial is consistent across several South African datasets.

In a recent compilation of returns to schooling estimates, Montenegro and Patrinos (2014) generate comparable simple OLS estimates of the Mincer equation for 139 countries. They argue that it is typically difficult to compare returns estimates across studies because of different sample definitions, different models employed to estimate returns, and differing choices of independent variables. As such, these authors compile harmonized household surveys and implement the same specification and estimation procedure across countries to generate comparable returns to schooling estimates. Their findings place South Africa as the country with the second highest returns to schooling (after Rwanda) out of the 139 countries, with a return to schooling of 21%.

2.3.2 Controlling for Bias in South African Studies

South African studies have attempted to address the sources of bias to some extent. Several studies have controlled for selection bias. Keswell (2004) and Keswell and Poswell (2004) estimate returns using both OLS and a Tobit approach. This accounts for censoring at zero, as wages are only observed for those who work. The Tobit approach allows estimation of both employment and earnings determination. Other authors employ a Heckman two-step model to explicitly control for selection into the employed sample. Serumaga-Zake and Naude (2003) use both a Heckman and a Double Hurdle model, where the latter estimates both the decision to participate in the labour force and the likelihood of employment. They find similar results from both models, which suggest a return to education of 11-12%.

Several studies have also attempted to correct for the bias due to the omission of the quality of schooling as a variable. Particularly if Mincerian equations are estimated by race, the differences in returns may be largely capturing the quality of educational institutions. Chamberlain and Van der Berg (2002) find that about half of labour market discrimination

in returns to education can be explained by the quality of the institution. Case and Yogo (1999) use national surveys of school quality, and find that a decrease in the pupil-teacher ratio of 5 students is correlated with an increase in the returns to education of 1 percentage point. Additionally, they find that school quality affects both the total years of completed education and the probability of employment.

However, the returns to schooling literature in South Africa has typically neglected the important sources of bias in the form of omitted variables. Keswell and Poswell (2004) argue that a lack of adequate data prevents scholars from comprehensively addressing these issues. However, complete neglect of these biases in the developing country context may be a major pitfall. As Lam and Schoeni (1993) argue, omitted variable bias is likely to be even more significant in developing countries than developed. This is on account of the stylized fact that family background appears to play a more important role in intergenerational mobility in developing countries. Hertz (2003) acts as the sole paper using South African data that makes a thorough attempt to deal with omitted variable bias.[1] He estimates the returns to education in South Africa using the 1993 PSLSD survey. He addresses omitted variable bias by using a within-family fixed effects differencing strategy. When accounting for this bias, he estimates returns to education of 5-6%. His OLS estimates of 11-13% were therefore biased upwards.

In sum, the South African research on returns to schooling tends to be largely descriptive. The literature has mainly studied how returns differ by race, and the shape of these returns over education levels. The work that has been done suggests that South Africa has a very high return to schooling compared to other countries. Some attempt in the literature has been made to account for selection bias. However, there is only one main paper, namely Hertz (2003), that makes a significant attempt to deal with omitted variable bias. There is thus significant room for research to be conducted that corrects for this endogeneity within the South African context, and thereby allows a more causal interpretation of the results.

3 Methodology

There are two novel features in the methodology of this paper compared to other returns to schooling studies in South Africa. The first is that it will make a thorough attempt to

[1] It must be noted that a recent paper has been published (Mariotti and Meinecke, 2014) which uses non-parametric techniques to deal with the endogeneity of education. However, these techniques are beyond the scope of this paper.

control for the endogeneity of the education variable that is caused by unobserved individual characteristics. The second feature will be the use of a panel dataset to do the analysis. This paper will represent the first attempt at using panel data to explore returns to schooling in South Africa. It will thus challenge the accepted wisdom that "the applicability of panel data to estimates of schooling returns is limited" (Harmon et al., 2000: 32).

3.1 Sample and Variable Selection

In this paper, the data is sourced from the National Income Dynamics Study (NIDS), which is the first national panel data study conducted in South Africa, comprising of 28,000 individuals and 7,200 households. The first wave took place in 2008, the second in 2010/2011, and the most recent wave in 2012/2013.

The sample in this study is made up of individuals aged between 16 and 65, who were not enrolled in schooling in any of the waves. This sample is then further reduced to those who have a regular job or some form of casual employment. This is in accordance with Hertz (2003), and therefore excludes self-employment, because it is subject to greater measurement error. Individual monthly income is thus calculated as the sum of regular employment income, plus casual income. The overall sample is 7185 individuals. It must be noted that this is an unbalanced sample in that it is not the same group of individuals in each cross-section. This is appropriate for the cross-sectional analysis which will consider the returns to education for each separate point in time. However, in the later panel analysis, this sample will be updated to a balanced panel of working individuals.

Both of the income variables (pay from main employment and casual earnings) are deflated before being aggregated. This is to ensure that income is comparable across years. Furthermore, fieldwork for NIDS took place over several months even within waves. For example, in the second wave, interviews were conducted between May 2010 and August 2011. Thus, deflating is necessary to ensure that the income figures represent equivalent amounts even within waves. Deflations are based on CPI figures,[2] with the base month being the modal month of wave 3, namely August 2012. If the month that the interview took place is missing, then it is assumed that this interview took place during the modal month for that particular wave.

The analysis that follows can be said to only consider returns to schooling for those individuals with paid employment. This paper therefore does not consider issues of bias around selecting

[2]Available on the Stats SA website, http://www.statssa.gov.za/keyindicators/CPI/CPIHistory.pdf

into the labour force and selecting into employment. This accords with the work of scholars such as Garcia et al. (2005) and Harmon et al. (2000) who make a similar assumption. As such, the results cannot be considered to be widely generalisable. They have no insight into the expected earnings given education for the general population in South Africa. Rather, the results are narrowly applicable to those who have both decided to participate in the labour force and have found work in the form of permanent or casual employment.

Several assumptions are made regarding the choices of variables for the Mincer equation. Firstly, the education question in the NIDS questionnaire is phrased as the highest grade or qualification that the respondent received. This variable is therefore converted into the linear variable: number of years of education. This recoding requires several assumptions about the typical number of years required to obtain a particular level of education.[3] This traditional variable of "years of schooling" allows for the cleanest interpretation of returns to schooling and will thus be used as the primary measure of education in the following estimates of the Mincer equation.

A secondary measure of education will also be used. Dummy variables for the completion of some level of primary, secondary, or tertiary qualification are created. Additionally, there is a dummy for whether the individual attained their Matric qualification, which will act as the base group in the analysis. There are several benefits of this secondary measure of education. It is likely to be subject to less measurement error because no assumptions had to be made regarding the number of years needed to obtain this qualification. Furthermore, it could capture nonlinear "sheepskin" effects in returns to schooling.

A further decision is to use age in place of experience. The NIDS survey does not have a direct question on experience, and it would be highly difficult to create a proxy for this variable with the large amount of casual work and unemployment that exists in South Africa. The choice of age accords with other South African papers, such as Keswell and Poswell (2005), who argue that the traditional proxy for experience is flawed in South Africa due to high grade repetitions and unemployment. As such, age is subject to the lowest measurement error and, as discussed above, the results only deviate slightly from an equation where experience

[3]The assumptions for coding education are as follows: A certificate = 1 year; A diploma = 2 years; A masters/doctorate = 7 years. If Grade 12 was not attained but the respondent had a certificate or diploma, it was assumed they had completed 10 years of school first. As such, a certificate with less than Grade 12 was assumed to be 11 years of schooling. A diploma with less than Grade 12 was coded as 12 years of schooling. A certificate with Grade 12 was assumed to be 13 years of schooling. A diploma with Grade 12 was assumed to be 14 years of schooling. A bachelors degree was coded as 15 years of schooling. A bachelors degree with a diploma was assumed to be 17 years of schooling. An Honors degree was assumed to be 16 years of schooling. A masters or doctorate was assumed to be 19 years of schooling. All NTC qualifications were assumed to be 13 years of education.

is used. A final benefit of using age is that it is an exogenous variable and is not subject to choices that would be determined by unobserved individual characteristics. Conversely, an experience variable would need to be instrumented for, and would therefore reduce the precision in the estimation of the Mincer equation (Garcia and Montuenga, 2005).

Other covariates that are added to the Mincer equation include racial dummies, a gender dummy, a control for if the individual is a member of a union, and a rural dummy. This last variable is coded so that respondents in tribal homelands or in formal rural areas are considered rural. The choice of these covariates follow other papers on returns to schooling in South Africa; see for example Keswell and Poswell (2005) and Hertz (2003). As such, the additional controls in this analysis make for an "augmented Mincer equation".

3.2 Descriptive Analysis

Table I presents the weighted descriptive statistics for each wave. The statistics have been weighted with a composite of design and post-stratification weights. The design weights take into account the unequal probabilities of including each enumerator area and household in the sample. These weights also compensate for household nonresponse. Post-stratification weights are then incorporated so as to conform the sample to the Stats SA Mid Year Population Estimates (De Villiers et al., 2013). The standard errors are also adjusted based on the strata and cluster sampling design. The cluster in wave 2 and 3 is assumed to be the initial cluster in wave 1. However, weighting within a panel context will be discussed further in the panel data analysis section.

Education increases slightly over the period, suggestive of misreporting as the sample attempted to exclude individuals enrolled in schooling during this period. The proportion of individuals with secondary or tertiary schooling similarly increased over this period. However, it must be noted that inclusion into the sample group is fluid and dependent on having work. This samples for each wave thus include temporary sample members and continuous sample members. As such, it is not the same group of individuals in each cross-section, and this unbalanced sample is part of the explanation for the change in education over the period.

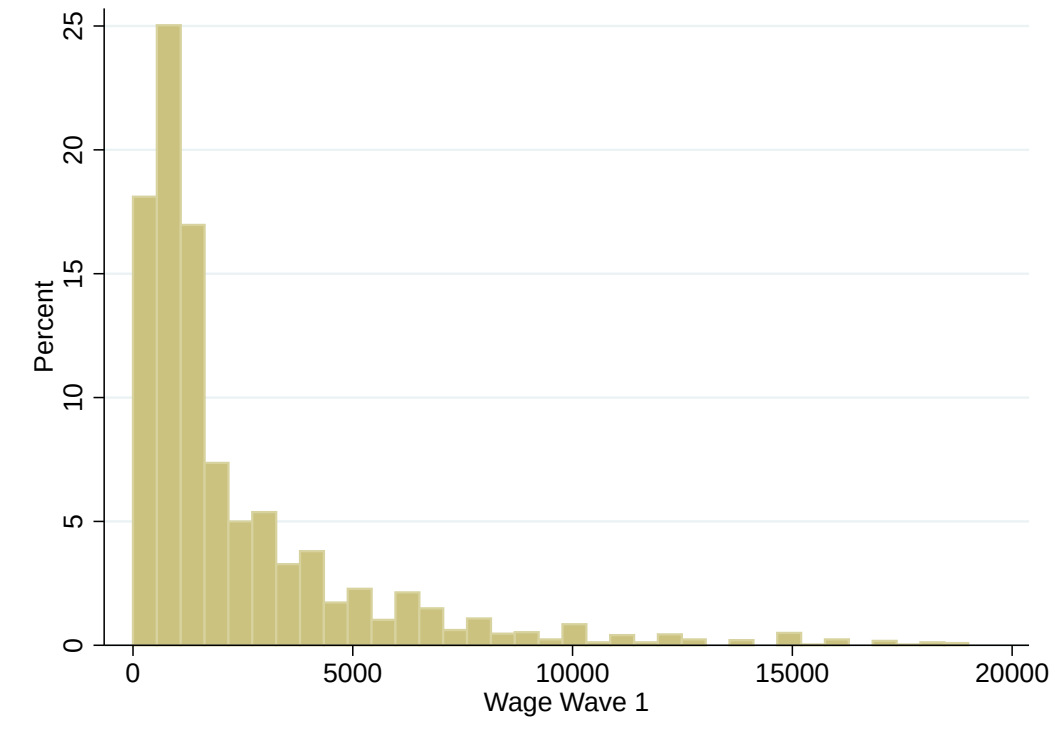
The average number of years of schooling in the sample is 10-11 years, with 32-36% of the sample having stopped their education during their high school years. Approximately 21-25% of the sample has finished Matric. The average age in the sample is 37 years in 2008 and 38 years in 2012. Black workers represent around 72% of the sample, with coloureds at 10%, and whites at 13%. This racial break-down of those with employment does not therefore accord

with the racial breakdown of the wider population in South Africa, and implies that whites still have greater access to job opportunities. Unionized workers represent approximately 30-33%, whilst rural workers make up 22-24% of the sample.

TABLE I: DESCRIPTIVE STATISTICS

Variable	Description	Wave 1	Wave 2	Wave 3
wage	Sum of deflated main and casual earnings	4999.491	6420.125	5574.998
		(398.886)	(758.347)	(415.679)
lwage	Log of Wages	7.978	8.086	8.127
		(0.060)	(0.062)	(0.052)
educ	Years of Schooling	10.051	10.632	10.765
		(0.202)	(0.210)	(0.164)
primary	=1 if Grade 7 or less	0.220	0.177	0.147
		(0.016)	(0.018)	(0.016)
secondary	=1 if Grade 8 - Grade 11	0.319	0.321	0.361
		(0.017)	(0.019)	(0.020)
matric	=1 if Grade 12 completed	0.251	0.237	0.212
		(0.015)	(0.016)	(0.013)
tertiary	=1 if tertiary qualification	0.210	0.265	0.281
		(0.022)	(0.023)	(0.021)
age	Age	37.114	38.349	38.183
		(0.385)	(0.460)	(0.450)
agesq	Age Squared	1487.773	1575.900	1566.041
		(29.749)	(36.995)	(37.119)
female	=1 if Female	0.425	0.433	0.429
		(0.019)	(0.020)	(0.019)
rural	=1 if lives in a tribal homeland or formal rural area	0.249	0.233	0.225
		(0.028)	(0.025)	(0.023)
black	=1 if Black	0.728	0.721	0.738
		(0.030)	(0.036)	(0.032)
coloured	=1 if Coloured	0.104	0.117	0.104
		(0.018)	(0.025)	(0.021)
asian	=1 if Asian	0.027	0.029	0.030
		(0.011)	(0.011)	(0.014)
white	=1 if White	0.141	0.133	0.128
		(0.026)	(0.027)	(0.022)
union	=1 if Member of a trade union	0.330	0.333	0.306
		(0.024)	(0.025)	(0.024)

Figure I: Histogram of Wages, Wave 1



In Figure I, a histogram of the distribution of wages is plotted for the sample for the first wave. It shows that 86% of individuals in the sample earn less than R5000 for wave 1. The largest spike of earned wages falls between R500 and R1000 per month. Figure II plots log wages against education. This gives an indication of the correlation between education and wages. For each wave, this correlation is positive, and slightly convex. However, the plot appears to be largely flat until the end of 12 years of education. This suggests that wages are not hugely different amongst those with incomplete schooling, and that a noticeable education premium only manifests itself after a Matric certificate is obtained.

Figure III considers educational attainment by race for the sample. Blacks make up three-quarters of those individuals whose maximum level of education is primary school. Conversely, there are no whites who have less than 7 years of education. The proportion of blacks in each education category decreases with each level, until only 59% of individuals with more than 12 years of schooling are blacks. This is despite the fact that blacks make up about 70% of the sample. Conversely, 24% of those with a tertiary qualification are whites. The pie chart that most closely reflects the overall sample racial break-down is that of secondary school. This suggests that access to secondary is the most universally attainable in South Africa, whilst a higher level of education is still biased toward whites.

Figure II: Plot of Monthly Wages against Education for each Wave

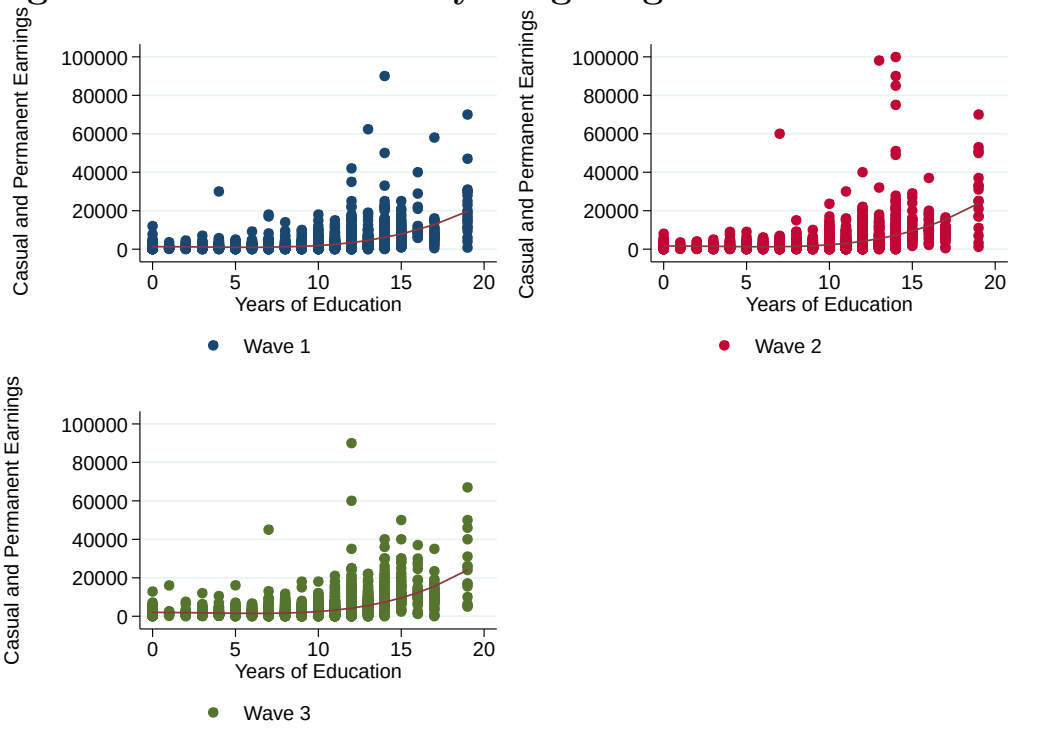


Figure III: Pie Charts of Education Attainment by Race

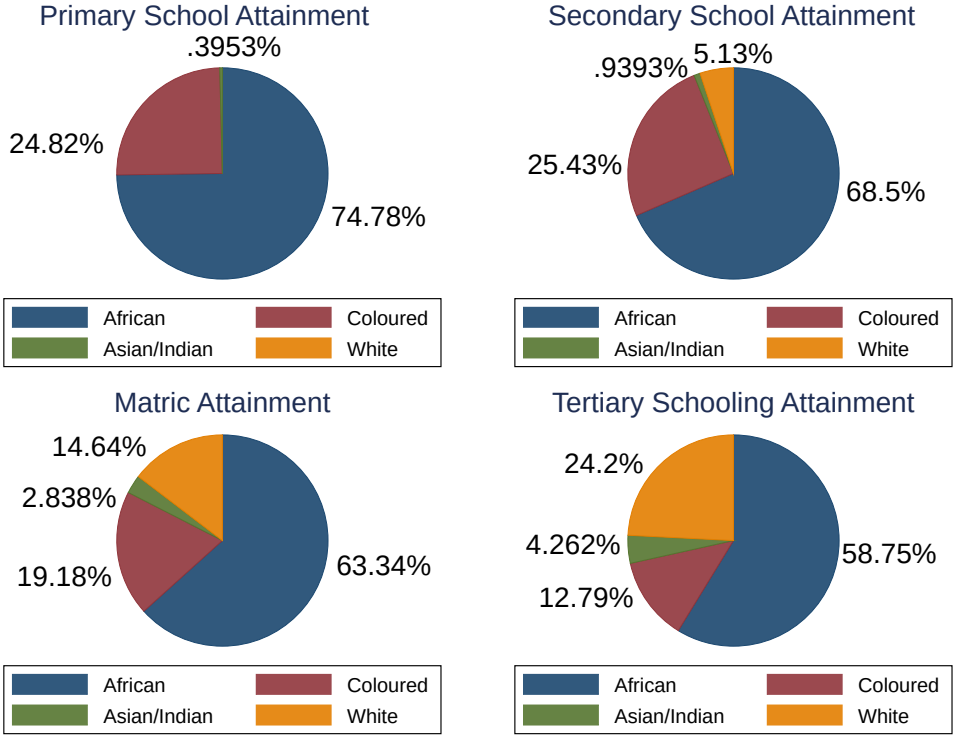


Figure IV: Bar Chart of Educational Attainment by Location and Gender

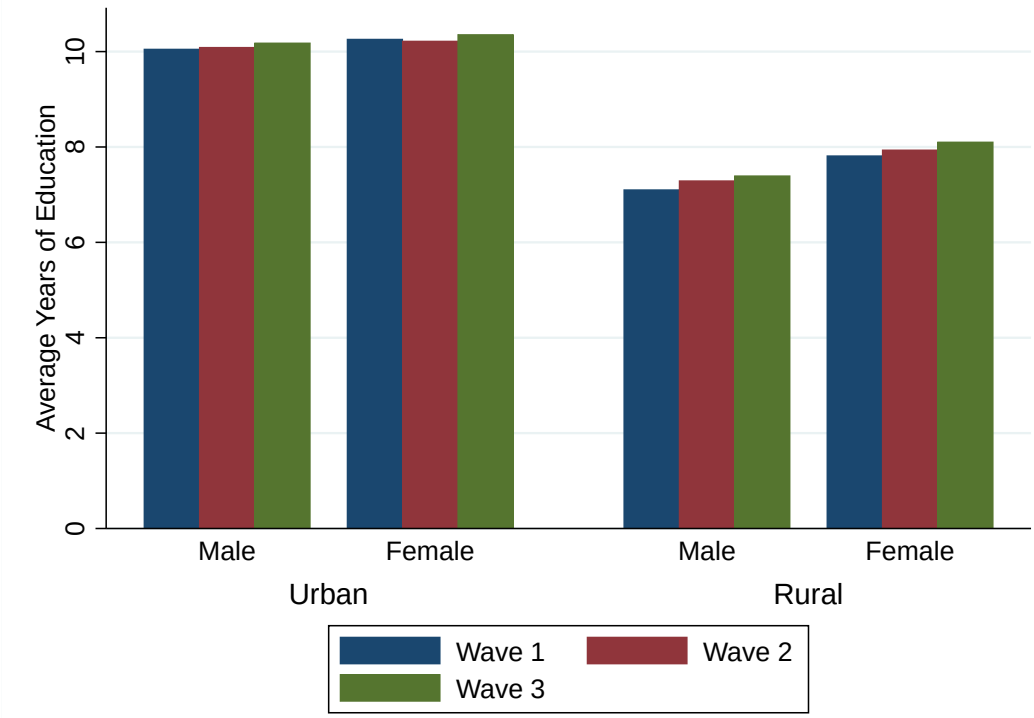
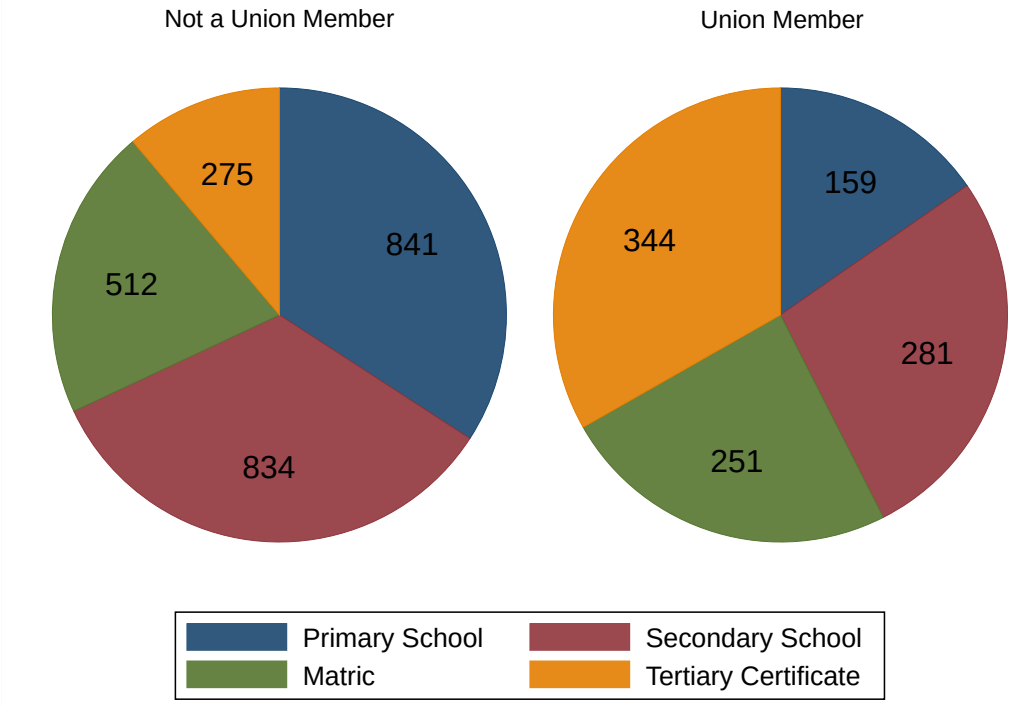


Figure IV considers the average educational attainment by location and gender. For both males and females, the average level of education is lower in rural areas than urban. This is expected and it is likely due to both demand-side factors (employers do not require as much skilled labour in rural areas) and supply factors (low access to schools). The average education between males and females is not hugely different, although females have a slightly higher level of education. This difference appears more stark in rural areas, and could be due to the fact that lower-skill farmworkers are perhaps more likely to be male, and these workers thus have a greater incentive to leave school earlier.

Finally, Figure V considers educational attainment by union membership. The labels on the slices represent the total number in that category for wave 1. The graph thus confirms Table I in that there is a smaller proportion of unionized workers in the South African labour market. Almost three-quarters of non-union members tend to have only primary or secondary schooling. The smallest group within non-unionized members are those who possess a tertiary qualification. Conversely, amongst union members, the majority have studied for more than 12 years. There are many possible reasons behind this result. One option is that a lack of education is acting as a barrier to union membership in South Africa, perhaps due to a lack of resources or skills to obtain information about unions. It also speaks to the nature

Figure V: Pie Chart of Educational Attainment by Union Membership for Wave 1



of jobs in South Africa, with perhaps fewer unions representing individuals doing casual or low-skilled work.

In summary, this section has conducted a descriptive analysis of the sample. The sample attempted to isolate working individuals who are no longer enrolled in any form of schooling. Amongst this sample, there is a positive correlation between education and income. However, no causal inferences can be made from this, as this correlation could be driven by union membership, race, or urban location. Furthermore, there could be unobserved characteristics driving this correlation. As such, a regression analysis which allows a "ceteris paribus" interpretation must be conducted before any causal conclusions can be made.

3.3 Missing Data

Missing income data may be a problem if certain subsets of the sample prefer not to divulge their earnings to enumerators. If the probability of income data being missing depends on education, then returns to schooling might not be an accurate reflection of the true returns to the working sample. The income data in NIDS is of a coarse nature in that there is a

mixture of observation types, where the observed values do not fully reflect the true values of the data (Daniels, 2008 and 2012). Some individuals reported exact income figures. However, others refused or did not know the exact amounts. In this case, they were prompted with fifteen income brackets. In the first wave of NIDS, this question was phrased using showcards, where the individual was required to identify the correct income bracket. In the second and third waves, this question changed to a "more than/equal to/less than" phrasing to try better elicit a response. The final observation type is missing income information. This is due to a continued lack of response in the form of refusal, a "don't know" response to the bracket question, or a survey collection error whereby the information is unaccountably absent.

For the income variables in NIDS wave 1 that used showcards of fifteen income brackets, the coarse data structure can be presented as follows (Daniels, 2012). Let $X = x_{ij}$ represent the complete data matrix. Let G represent the coarsening variable which determines the process by which x is mapped into $x_{observed}$. Coarse data is then represented in the following framework:

$$x_{observed,ij} = \begin{cases} \{x_{ij}\} & if \quad G_{ij} = \{0\} \\ [x_L \leq x_{ij} < x_U) & if \quad G_{ij} = \{1, 2 \ldots 15\} \\ \Psi & if \quad G_{ij} = \{16, 17, 18\} \end{cases} \tag{11}$$

When G_{ij} is equal to 0, then $x_{observed}$ is the set of exact responses in income. When G_{ij} is between 1 and 15, the observed data is contained within any of the fifteen income brackets, where x_L is the lower bound and x_U is the upper bound. Finally, $x_{observed}$ falls within the sample space of X, denoted as Ψ, when the response is "don't know" or a refusal or it is missing for other reasons. In this latter case, G_{ij} is denoted as 16, 17 or 18. As such, coarse income data consists of continuous, bounded and item missing responses.

The mixture of these data types for income is analyzed in Table III. The two types of employment that are considered in this paper (main and casual) are presented for each wave. As can be seen, the proportion of exact responses increased over the waves. For example, with main employment, the proportion of exact responses increased from 81% to 88%, and then to 96% in the final wave. Concurrently, the proportion of refusals fell drastically – for main employment it fell from 10% to 1% over the waves. This result is expected in a panel study, as respondents grow more trusting of the survey team and the confidentiality of their responses over time. They realise that there is no risk of disclosure of their personal information.

TABLE II: COARSE DATA STRUCTURE

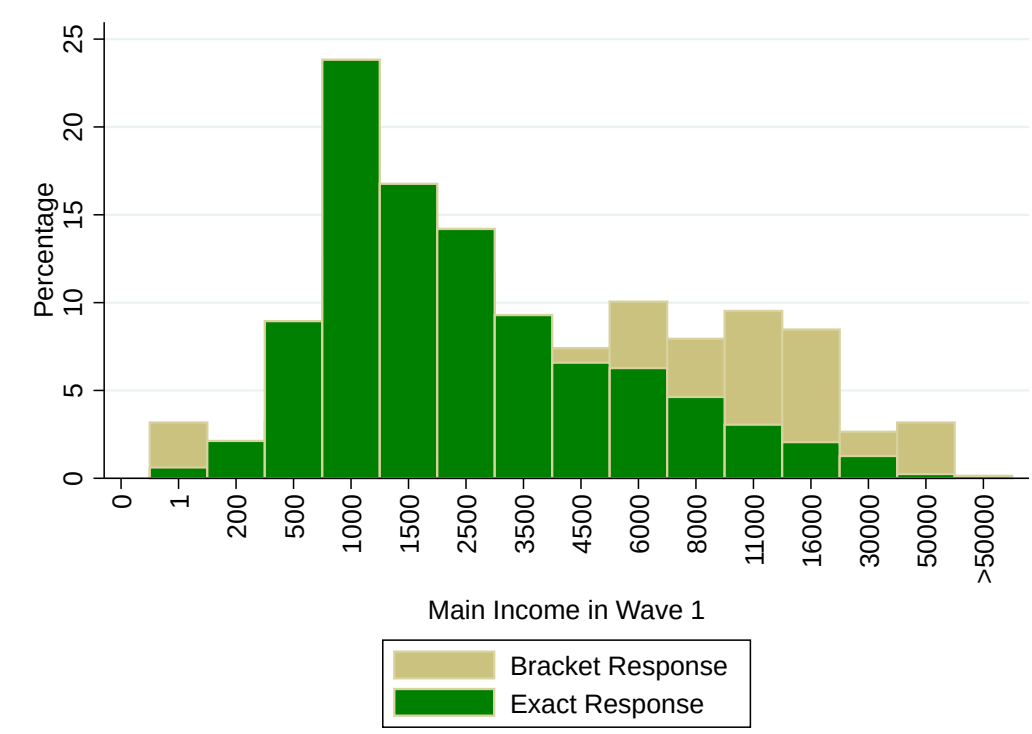
Wave	Variable	Exact	Brackets	Don't Know	Refuse	Missing	Total
1	Main Pay	2920	189	63	356	71	3599
		81.13%	5.25%	1.75%	9.89%	1.97%	100%
	Casual Pay	585	20	15	36	12	668
		87.57%	2.99%	2.25%	5.39%	1.80%	100%
2	Main Pay	2904	236	9	91	48	3288
		88.32%	7.18%	0.27%	2.77%	1.46%	100%
	Casual Pay	393	24	1	3	0	421
		93.35%	5.7%	0.24%	0.71%	0%	100%
3	Main Pay	3380	99	6	44	7	3536
		95.59%	2.8%	0.17%	1.24%	0.2%	100%
	Casual Pay	466	11	1	2	4	484
		96.28%	2.27%	0.21%	0.41%	0.83%	100%

Another interesting point to note is that the proportions of "don't know" responses fell over time, suggesting that this was used as an excuse to refuse answering in early waves. Finally, the proportion of bracketed responses increased in wave 2 from wave 1, suggesting that the alternative phrasing of the bracket question was perhaps less intrusive than the showcard approach. However, these proportions dropped again in wave 3, as proportionately more individuals were willing to give exact responses.

The variable that presents the most concern is that of main pay in wave 1. Only 81% of the responses are exact figures, and there are almost 700 individuals that would be excluded from the analysis if the coarse data structure in this variable is ignored. Furthermore, Figure II compares the percentage of individuals for each income category giving exact responses against the percentage answering in brackets for wave 1. As can be seen, a greater relative percentage of individuals who give bracketed responses are in the higher income categories. Conversely, exact responses are proportionately higher for the income brackets below R3500 per month. Higher earning individuals thus appear to be less willing to disclose their earnings. This is in accordance with Casale and Posel (2005) who find a significant difference between individuals who give exact income responses and those who defer to a bracket response.

Overall, this analysis has suggested the need to take the coarse data structure into account. In particular, for main income in wave 1, there is a high proportion of bracketed and missing data. It further appears that higher income individuals are less willing to give exact income figures. Failing to account for bracketed and missing income data may therefore bias the

Figure VI: Histogram of Main Wages Comparing Bracket and Exact Responses



returns to schooling estimates in a downward direction.

3.4 Estimation Procedure

The analysis in this paper will comprise of two sections. The first part will be the cross-sectional analysis of returns to schooling. This will provide initial benchmark results for returns to schooling where no attempt to address endogeneity is made. The second section will estimate the returns to education using panel data. A more detailed break-down of the steps that will be taken in each section will now be given.

In the cross-sectional analysis, simple OLS estimation of the augmented Mincer equation will be presented. This estimation will be conducted by treating each of the three waves as independent surveys. Such an approach will allow some consideration of changes in the variables over time. Thereafter, a univariate multiple imputation method will be implemented for the income variable in wave 1. As discussed previously, this variable is subject to a large proportion of missing data and bracketed responses. Multiple imputation is a Monte-Carlo technique where missing values are imputed several times conditional on observed data. This

is superior to single imputation, in that it incorporates random variation in determining the parameter estimates (Allison, 2001).

In the second section, the data will be considered in longitudinal form. The Mincer equation will be estimated using the differencing approaches, namely first-differencing and fixed-effects estimation. The results of this analysis are unlikely to be significant or to be in accordance with expectation. This is because the education coefficient will only be identified for individuals who have changed their education. It is unlikely that this subgroup will be representative because the sample was intended to capture individuals no longer enrolled in education.

As such, the analysis will then move on to the Hausman-Taylor (HT) approach. A starting point will be to create a balanced panel, where there are the same number of time-series observations for each individual. In this case, it implies that the sample will comprise of only those individuals with permanent or casual employment across all three waves. The HT approach will allow both the identification of the education coefficient and deal with its endogeneity. The HT model is set up as follows,

$$ln(y_{it}) = X_{it}\beta + Z_i\gamma + c_i + u_{it} \qquad t = 1, ..., T; \;\; i = 1, ..., N \qquad (12)$$

where X_{it} refers to the time-varying regressors, Z_i the time-invariant regressors, c_i the individual time-constant heterogeneity, and u_{it} the idiosyncratic error term. X and Z can be partitioned further, $X_{it} = [X_{1it}, X_{2it}]$ and $Z = [Z_{1i}, Z_{2i}]$, with X_{1it} and Z_{1i} assumed to be exogenous, but X_{2it} and Z_{2i} are correlated with c_i. Education forms part of Z_{2i} in the modeling of returns to schooling as it is time-constant and endogenous. All regressors are assumed to be orthogonal to u_{it}.

In the first stage of the HT approach, a fixed effects regression is conducted on a model that contains only the time-varying regressors. The resulting estimates are consistent because the bias-causing individual effects are removed. Residuals can then be calculated and used to derive estimates of $\hat{\Omega}$, the variance-covariance matrix of the composite error term, $v_{it} = c_i + u_{it}$. The model can then be weighted through premultiplication by $\hat{\Omega^{-1/2}}$. This generalized least squares transformation ensures efficient estimates.

In the second stage of the HT approach, an instrumental variables estimation is conducted using internal instruments. The endogenous time-invariant regressors, Z_{2i}, are instrumented for with the individual means of the time-varying exogenous variables, $\bar{X_{1i}}$. This instrumental choice meets the two conditions for instrumentation. Firstly, $\bar{X_{1i}}$ is exogenous to both c_i and

u_{it} by assumption, therefore fulfilling the instrument exogeneity requirement. Secondly, the exogeneity of $\bar{X_{1i}}$ to c_i and u_{it} necessarily implies that $E[\bar{X_{1i}}Z_{2i}] \neq 0$ in order to make the between-regression hold. The instrument relevancy condition (of correlation between the instrument and the endogenous variable) therefore holds (Wooldridge, 2002).

Furthermore, any endogenous time-varying regressors can be instrumented for with their demeaned values: $\ddot{X_{2it}} = X_{2it} - \bar{X_{2i}}$. These instruments obviously meet the relevance condition, as the demeaned values are necessarily correlated with the level values. The fulfillment of the exogeneity condition is determined through linear projections: $X_{2it} = \phi c_i + u_{it}$ and $\bar{X}_{2i} = \bar{\phi c_i} + \bar{u_i}$. Subtracting these two linear projections to get a demeaned linear projection yields $\ddot{X}_{2i} = u_{it} - \bar{u}_i$, which shows that $\ddot{X}_{2i}$ is unrelated to c_i, despite the fact that X_{2it} is correlated with c_i. The demeaned values of X_{2it} thereby meet the instrument exogeneity condition.

The instrument set can thus be summarized as follows, where the identifying requirement is that there are at least as many time-varying exogenous regressors as there are endogenous time-invariant ones:

$$[X_{1it} \quad \ddot{X}_{2i} \quad Z_{1i} \quad \bar{X}_{1i}] \qquad (13)$$

In the analysis that follows, the HT results will be compared against the results of a random effects estimation. As discussed previously, a random-effects approach represents one extreme where all the regressors are treated as exogenous. These results are likely to be biased, and it will thus be of interest to see how the results change when education is considered to be an endogenous variable.

In summary, this section has set out the sample and procedure that will be followed in this analysis. The first stage of cross-sectional analysis will reflect the typical South African study on returns to schooling. However, as has been argued, when using OLS estimation, a causal interpretation is unconvincing. As such, the second section of this paper will form the contribution to the literature. It will use a panel dataset and an internal instrumentation technique to control for endogeneity in education. This procedure has not yet been applied to South African data.

4 Results

4.1 Cross-sectional Analysis

The Mincer equation is estimated using OLS and these initial results are presented in Table III. Each wave is considered separately, and returns are estimated cross-sectionally. The standard Mincer equation has been augmented with racial controls, and dummies for urbanization and unionization. The education coefficient is highly significant for each of the waves. In wave 1, the estimates suggest that an additional year of education is expected to increase wages by 10.9% on average, holding all else constant. In wave 2, the returns to schooling increase to 12%, whilst wave 3 has the highest expected returns at 14.0%.

However, when industry and occupation dummies are added to the regressions, the returns to schooling drop by approximately four percentage points. This suggests that wages in South Africa are significantly determined by the sector and type of work, and failing to control for these leads to an overestimate of the true effect of education on wages. This result therefore stresses the importance of the inclusion of these dummies. Other points of interest include the slight inverted u-shaped relationship that wages have with age. An additional year of age increases earnings, but at a decreasing rate. The quadratic term in age does not appear to be individually significant in wave 1 and wave 3. However, the age controls are together jointly significant for all waves.

The female coefficient is highly significant for each wave, and suggests that females earn 33-38% lower than males on average, when holding all else constant. Blacks and coloureds are expected to earn significantly less than whites, a factor that is expected given the history of South Africa. There are no significant differences in wages between Asians and whites. Individuals living in rural areas earn significantly less than urban areas. However, this differential appears to have fallen, as it was at 23% for the first wave, and reduced to only 13% in the third wave. There is also a significant union premium in operation in the South African labour market, but it has similarly declined, from 30% in the first wave, to 27% in wave 3.

4.2 Multiple Imputation of Missing Income Data

There are several approaches to account for coarse data. Regarding interval data, a midpoint of the interval can be used. However, this approach can lead to bias, and will create spikes in

TABLE III: INITIAL OLS REGRESSIONS

	Wave 1	Wave 1	Wave 2	Wave 2	Wave 3	Wave 3
educ	0.109***	0.068***	0.120***	0.082***	0.140***	0.093***
	(0.009)	(0.007)	(0.010)	(0.010)	(0.009)	(0.009)
age	0.042	0.034	0.072***	0.069***	0.036*	0.046***
	(0.024)	(0.020)	(0.018)	(0.015)	(0.014)	(0.013)
agesq	-0.000	-0.000	-0.001**	-0.001***	-0.000	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
female	-0.421***	-0.332***	-0.406***	-0.385***	-0.381***	-0.337***
	(0.046)	(0.045)	(0.047)	(0.052)	(0.041)	(0.049)
black	-0.816***	-0.618***	-0.926***	-0.692***	-0.703***	-0.553***
	(0.095)	(0.078)	(0.103)	(0.103)	(0.097)	(0.092)
coloured	-0.642***	-0.439***	-0.624***	-0.458***	-0.599***	-0.460***
	(0.127)	(0.095)	(0.147)	(0.122)	(0.119)	(0.103)
asian	-0.323	-0.297	-0.013	-0.161	0.031	-0.023
	(0.168)	(0.156)	(0.162)	(0.151)	(0.261)	(0.177)
rural	-0.275***	-0.235***	-0.264***	-0.263***	-0.110*	-0.133**
	(0.054)	(0.053)	(0.066)	(0.060)	(0.056)	(0.049)
union	0.447***	0.302***	0.351***	0.234***	0.379***	0.270***
	(0.054)	(0.053)	(0.067)	(0.065)	(0.056)	(0.054)
Constant	6.362***	6.609***	5.873***	5.943***	6.201***	6.302***
	(0.443)	(0.394)	(0.386)	(0.345)	(0.289)	(0.298)
Industry&Occupation Dummies	NO	YES	NO	YES	NO	YES
Observations	15863	15638	13145	13015	13539	13351

Marginal effects; Standard errors in parentheses

Source: NIDS, weighted

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

the data (Wittenberg, 2008). Alternatively, a distribution such as the uniform distribution can be applied to the interval. This is also likely to be inadequate because it fails to take into account the curvature of probability in response - uniform distributions treat all points in the interval as equally likely (Daniels, 2008). Furthermore, regardless of the pre-specified distribution, a single deterministic imputation will always underestimate the variances because the uncertainty of this imputation is not fully taken into account (Allison, 2001).

A multiple imputation approach, as first developed by Rubin (1978 and 1987), deals with the above problems. This is a simulation-based approach, where missing data are predicted m times based on an imputation model, where m represents the number of imputations. This yields m simulated values, and thus m simulated datasets. When the analytic model is then implemented, statistics are estimated separately for each of the m datasets before being combined to produce the overall estimates and standard errors. These estimates are combined using Rubin's rules (1987), which creates errors by combining the within and between-imputation variation. This latter component would not be present in a single imputation, and thus multiple imputation adjusts the standard errors upwards. This better reflects the uncertainty inherent within the imputation process and captures the fact that the imputed values are not of the same quality as point values (Wittenberg, 2008).

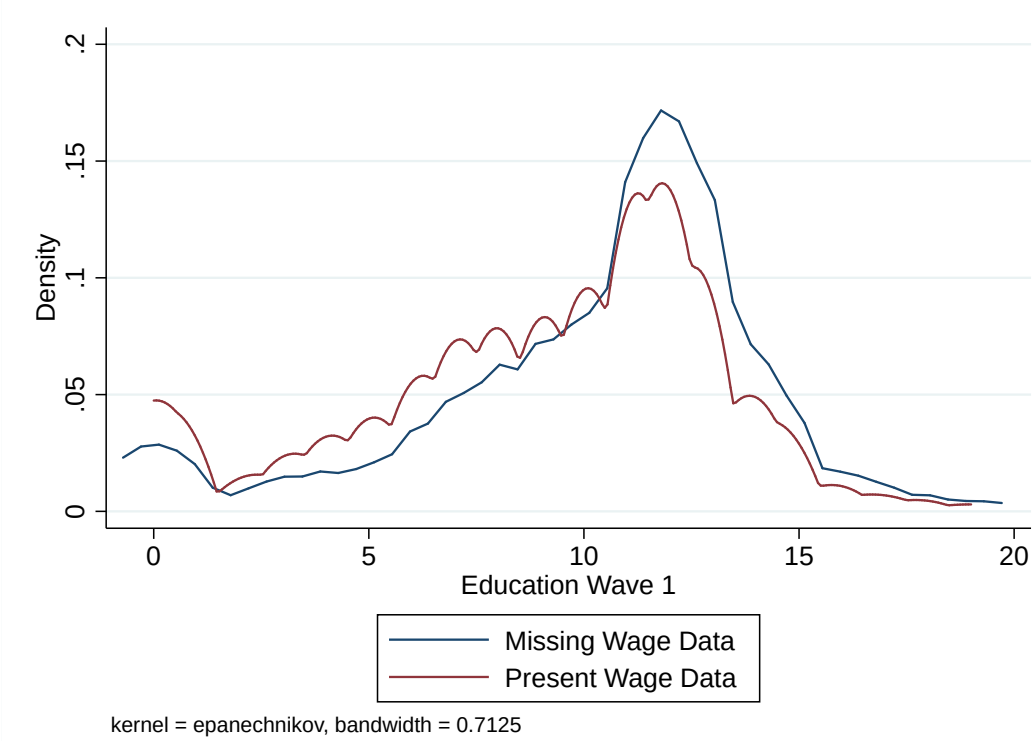
In this paper, the imputation will be based on an interval regression. Interval regressions are used when the data are censored into categories, yet the underlying income variable is continuous, as is the case with the bracket responses. The key assumption of the log-likelihood function applied to this imputation is that the log of income approximates a normal distribution (Daniels, 2008). For the bracketed data, this approach enables plausible draws within each interval, following a truncated normal distribution (Daniels, 2012). Missing values are imputed to any value following a continuous normal distribution (Vermaak, 2010).

An important assumption for conducting an imputation is that the data is missing (or coarse) at random (MAR). This means that the probability of income data being missing (or bounded) depends on observed covariates, but is unrelated to the true income value (see Rubin, 1976 and 1987). Using the set-up in (11), this assumption can be represented as follows,

$$f(M|X,\phi) = f(M|X_{observed},\phi) \qquad (14)$$

where M represents the missing data matrix, and ϕ symbolizes other unknown parameters. Missing data is thus fully predictable from the other observed variables. In this case, the missing data can be said to be ignorable, in that there is no need to model the process which

Figure VII: Kernel Density Plot of Education, Wave 1

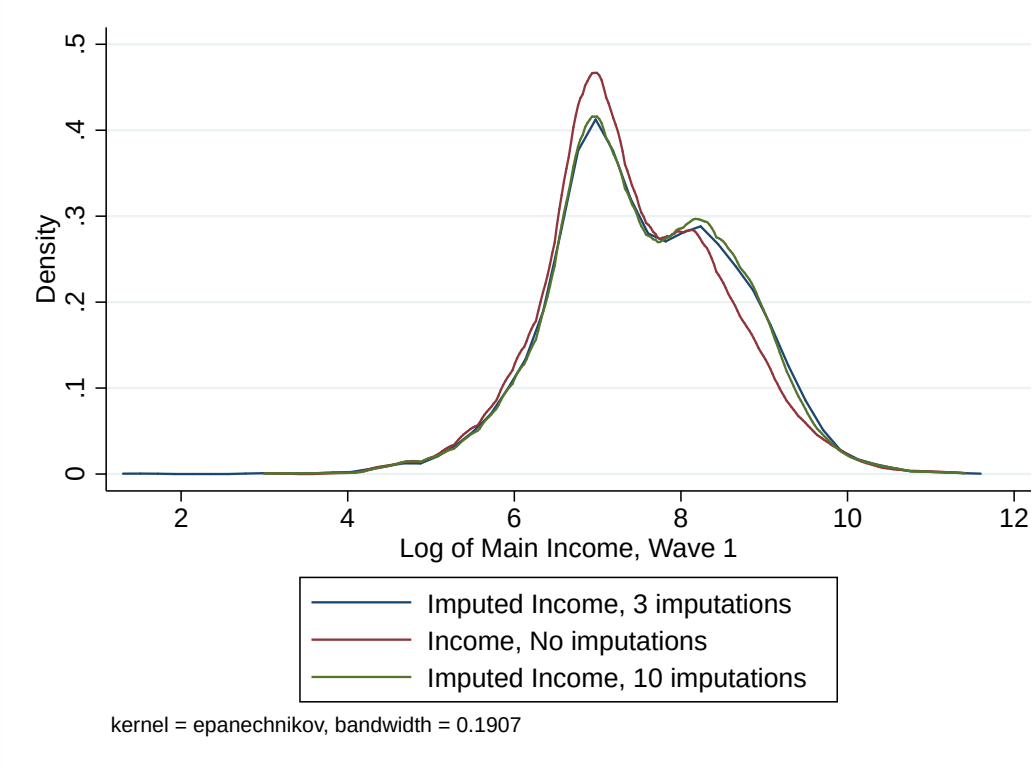


determines non-response. In other words, inferences can be made without knowing the nature of the mechanism that is driving missingness (Vermaak, 2010). However, ignorability does not imply that the results are unbiased; rather, the observed variables can be used to account for missing data without modelling the underlying non-response mechanism. Imputation thus makes a MAR assumption because the inference of the unknown parameters are assumed to be unrelated to the distribution of missingness (Lacerda et al,. 2007).

It is difficult to test for the MAR assumption. One approach is to follow Keswell and Poswell (2004) by considering the relationship of this missing data with education for wave 1. A kernel density plot of education is compared for individuals with missing wage data against those with accompanying wage data. A visual inspection of Figure VII suggests that there are some differences between the two plots. A greater proportion of individuals with secondary schooling and tertiary schooling have missing income data. Conversely, for lower education levels this pattern seems to reverse, and those with less than ten years of education seem more willing to divulge their earnings.

The correlation with education suggests that the income data is missing at random (MAR), in that the probability of income data being missing depends on observed variables, but is unrelated to the true income value (Daniels, 2012). A missing at random assumption implies

Figure VIII: Density of Imputed Main Income



that other variables in the dataset can be exploited to predict missingness, and can thus be used in a model to impute values for the missing income data. This suggests that the analysis in the paper could attempt to address and correct for missing income data using an imputation method.

As such, a multiple imputation method is employed for the main income variable in wave 1. This imputation is univariate in nature, because solely the coarse main income variable in the first wave is imputed for. Following Lacerda et al. (2007), the imputation model is similar to the analytic model to ensure the relationships between the missing values and the observed values are retained. The covariates in the imputation model therefore include an education variable, racial dummies, a gender and location dummy, and finally a unionization dummy.

Figure VIII plots the densities of income with and without imputations. The imputations have been conducted with m equal to 3 and 10. When income is not imputed for, it shows a higher peak in middle of the distribution. With imputations, the distribution shifts slightly to the right, as the values that are imputed tend to be for higher earning individuals. The plots with m as 3 or 10 are very similar and therefore suggest that the imputation has stabilised from a low number of repetitions. Overall, with visual inspection, there does not appear to

be a huge shift in the income distribution after imputation.

Table IV compares the wave 1 cross-sectional OLS regressions with and without imputations. The second column records the OLS results when m is 3, and it thus reflects the aggregation results of the 3 simulated datasets. The return to schooling increases slightly from 7.2% to 7.7% when main income is imputed. Other covariates similarly adjust slightly, but the level of significance does not change between the two specifications. As such, it appears that imputation has not made a large change to the results. Missing data and the coarse data structure will thus be ignored for the remainder of the paper. This will add to the ease of estimation in a panel data context, as combining a multiple imputation approach with the Hausman-Taylor estimation method would create analytical complexity that is beyond the scope of this paper.

TABLE IV: OLS RESULTS WHEN IMPUTING MAIN PAY, WAVE 1

	Without Imputations	With Imputations
educ	0.072***	0.077***
	(0.009)	(0.011)
black	-0.634***	-0.592***
	(0.107)	(0.112)
coloured	-0.437***	-0.405***
	(0.113)	(0.112)
asian	-0.194	0.022
	(0.172)	(0.299)
rural	-0.272***	-0.280***
	(0.057)	(0.064)
female	-0.288***	-0.297***
	(0.056)	(0.056)
age	0.087***	0.072***
	(0.019)	(0.020)
agesq	-0.001***	-0.001**
	(0.000)	(0.000)
union	0.289***	0.263***
	(0.062)	(0.059)
Observations	12963	.

Marginal effects; Standard errors in parentheses. Results are not weighted.

Source: NIDS

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In summary, a cross-sectional analysis has suggested that returns to schooling are between 7%-9% for the working sample in South Africa. This result has been verified when including imputations for bracketed and missing income data, with the returns to schooling remaining

within this band. However, the OLS specifications that have been conducted thus far are likely to be inappropriate for two reasons. Firstly, the OLS estimates are likely to be biased and inconsistent. As discussed previously, unobserved ability and family background effects are correlated with education, and education is thus an endogenous variable.

Secondly, the OLS standard errors are unreliable even if the unobserved variables are uncorrelated with the explanatory variables. The omitted variables of ability and family background are likely to be constant over time. As such, the very presence of a time-invariant effect will lead to correlated errors across time. The errors cannot thus be assumed to be independently distributed, and OLS is then not the most efficient estimation approach. In the following section, an attempt to deal with these two problems of OLS will be made using panel data.

4.3 Panel Analysis

A starting point of a panel data analysis is estimation using the main panel data techniques of time-differencing. As such, a fixed-effects and first-difference estimation approach is taken in Table V. These estimation strategies ensure that the individual, time-constant effects disappear. The former approach does this by subtracting the average of the variable over time for each individual, whilst the latter subtracts the lag of the variable. One downside to these estimators is that they cannot estimate time-constant variables, and the race and gender variables are thus omitted. However, education is the variable of interest, and it contains some variation over time and can thus be estimated.

A brief discussion on survey design adjustments and survey weights within a panel data context must first be conducted. These adjustments attempt to ensure that the results can be generalized to a population of interest. However, with a panel, it is no longer clear as to who this population is. The results cannot be generalized to the national population because the specific sample that is tracked neglects new settlements and over-samples particular groups over time.

Furthermore, with migration over the years, it is extremely difficult to decide on the nature of the survey adjustment. As Wittenberg (2013) points out, an adjustment for the primary sampling unit (cluster) is used to account for that fact that individuals tend to be more similar within the cluster, and this adjustment thereby prevents an overestimate of the standard errors. However, over time, individuals migrate to different areas. It is not clear if these common influences on individuals should be linked to the original cluster, the current clus-

ter, or the household. It comes down to the researcher's belief about the underlying social processes at work.

Finally, generalized least square (GLS) procedures, such as a random effects and the HT estimator, assume that the error structure is known. In a survey adjustment, the underlying assumption is that the error structure is incorrect, and weights are then used to correct for this. As such, statistical programs such as Stata do not allow for weighting with a GLS approach. In sum, the above discussion highlights the difficulty in using weights and cluster corrections in a panel data context. The analysis that follows therefore does not account for survey design or weights. The results are only applicable to the specific sample of analysis. However, since no attempt is made to correct for sample selection into employment in this paper, it implies that even without weights, the results are necessarily limited in generalisability.

Table V presents the results of time-differencing after the dataset has been converted into a longitudinal format. The results for the fixed effects and first differencing are counter-intuitive. The education coefficient is no longer significant and is actually negative for both specifications. The union variable is also negative for both specifications, which contradicts typical economic theory. The linear term of age and the rural dummy are the only explanatory variables that are significant in a panel-data context. The racial dummies have been eliminated from this model as they are time-invariant.

There are likely to be several reasons behind these strange results. Firstly, differencing and fixed effects methods rely on variation in the explanatory variables over the waves to identify the coefficient. However, the sample in the study has deliberately tried to isolate individuals who were not enrolled in school during this period. Any variation in education that is present is suggestive of misreporting or recall error. Secondly, the sample size of those whose education varied is also small given the nature of the created sample, and thus unusual results can be expected given this small n.

Time-differencing approaches are thus not appropriate for estimating returns to schooling. The Hausman-Taylor method is the one approach that deals with the problems encountered in using panel data for this analysis. Its generalized least squares approach adjusts the error term, whilst its use of internal instruments allows identification of time-invariant regressors that are endogenous. The Stata 12.0 version includes a command for the HT procedure and is used to estimate the following results.

TABLE V: PANEL REGRESSIONS

	Fixed Effects	First Differencing
educ	-0.0142	-0.0532
	(-0.53)	(-1.50)
age	0.0828**	0.0477
	(2.77)	(0.91)
agesq	-0.0002	-0.0002
	(-0.80)	(-0.37)
rural	-0.278*	-0.359*
	(-2.00)	(-2.15)
union	-0.0134	-0.0484
	(-0.30)	(-0.90)
constant	5.436***	0.0635
	(7.95)	(0.90)
Number of Observations	7202	2524

Marginal effects; Standard errors in parentheses

Source: NIDS

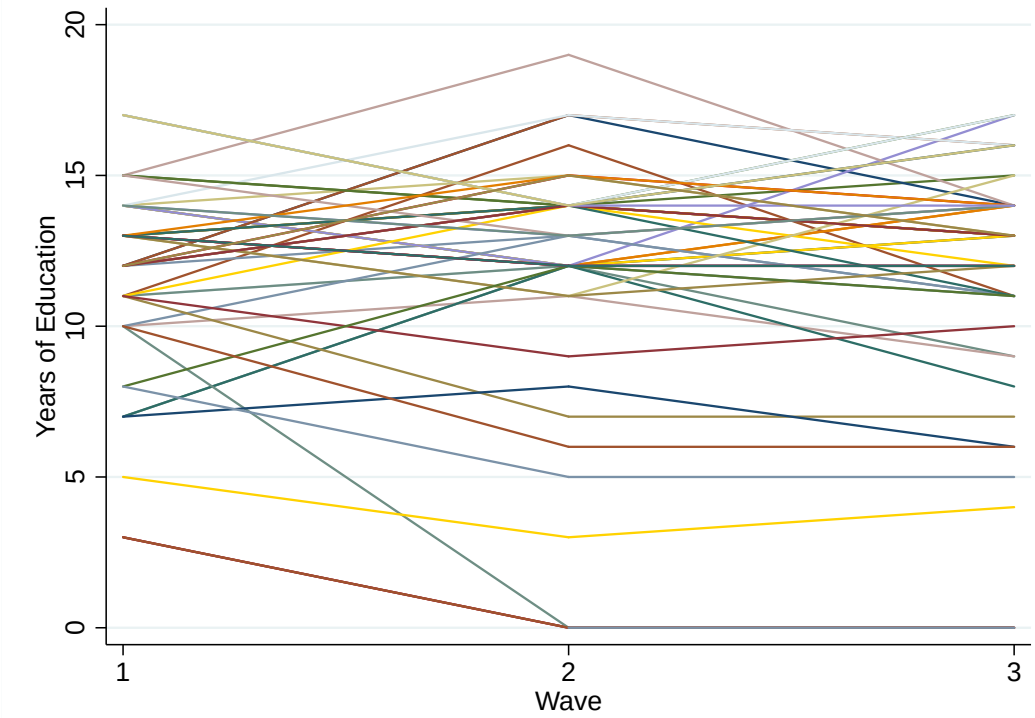
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3.1 Creating a Balanced Panel Dataset

A balanced panel refers to the case where there is information for all n cross-sectional units, over all the time periods under consideration. In the analysis that follows, the sample is thus restricted to a balanced panel with workers having regular or casual employment for all three waves. This ensures that labour market information is available for each of the individuals across the three waves. This leads to an overall sample of 1141 workers. This sample is thus entirely composed on continuous sample members.

However, 204 individuals in this sample had education that changed over these waves, despite reporting that they were not enrolled in education over this time period. Additionally, 57 individuals even had education that decreased over the waves. This is shown visually in Figure IX. As can be seen, there are various patterns in the reporting of education, and all are illogical due to a decrease in education at some point. One individual reports going from 10 years of education to 0 years. Another individual increased his level of education from 12 to 17 years, despite there only being a 2 year gap between the waves. This suggests that education changes are likely to be a function of recall error, or deliberate misreporting. Including these individuals will distort the results and will not allow an accurate reflection of returns to schooling. As such, the sample is further reduced to the 937 who did not change

Figure IX: Years of Schooling of Individuals who Reported a Decrease in Education



their education over the three waves.

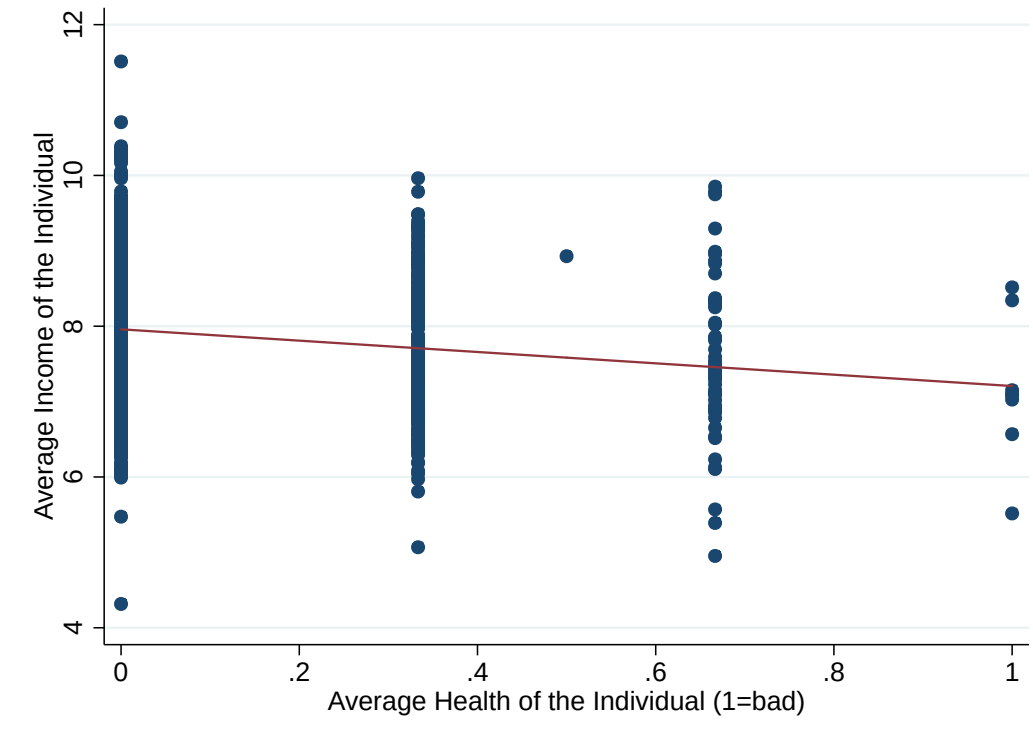
The descriptive statistics for this narrow sample are presented in Table VI. Despite the sample now including solely those individuals consistently employed over all three waves, the descriptive statistics do not change hugely from the cross-sectional samples Table I. Wages and educational levels appear to be similar between the two tables. One noticeable difference is that the balanced panel sample appears to be older than the cross-sectional sample (39-43 years old, compared to 37-38 years in Table I). This suggests that perhaps age is a determining factor in job retention and stability of work opportunities. A further significant difference is that the balanced panel contains a larger proportion of unionized workers. Again, this could imply that a key role of unions is to ensure job retention.

A new variable that is added to the balanced panel covariates is that of health. This is in accordance with Hausman and Taylor (1981) who use a measure of health as a time-varying exogenous instrument for education. The health variable in NIDS is derived from a perceived self-reported measure. Individuals are coded as having bad health if they report that their health is "fair" or "poor", whilst they are coded as being in good health if their reported health is "excellent", "very good" or "good". Figure X shows a positive correlation

TABLE VI: BALANCED PANEL DESCRIPTIVE STATISTICS

Variable	Description	Wave 1	Wave 2	Wave 3
lwage	Log of Wages	7.573	7.853	8.100
		(0.981)	(1.103)	(0.971)
educ	Years of Schooling	9.259	9.506	9.709
		(4.158)	(4.245)	(4.264)
primary	=1 if Grade 7 or less	0.274	0.275	0.271
		(0.446)	(0.446)	(0.445)
secondary	=1 if Grade 8 - Grade 11	0.314	0.304	0.304
		(0.464)	(0.460)	(0.460)
matric	=1 if Grade 12 completed	0.213	0.186	0.153
		(0.410)	(0.389)	(0.360)
tertiary	=1 if tertiary qualification	0.196	0.234	0.270
		(0.397)	(0.423)	(0.444)
age	Age	39.148	41.748	43.536
		(9.439)	(9.346)	(9.370)
agesq	Age Squared	1621.568	1830.202	1983.181
		(745.379)	(781.840)	(819.210)
female	=1 if Female	0.490	0.496	0.497
		(0.500)	(0.500)	(0.500)
rural	=1 if lives in a tribal homeland or formal rural area	0.379	0.379	0.382
		(0.485)	(0.485)	(0.486)
black	=1 if Black	0.719	0.702	0.713
		(0.449)	(0.457)	(0.452)
coloured	=1 if Coloured	0.216	0.230	0.220
		(0.412)	(0.421)	(0.414)
asian	=1 if Asian	0.011	0.010	0.012
		(0.104)	(0.104)	(0.112)
white	=1 if White	0.053	0.055	0.053
		(0.224)	(0.229)	(0.224)
union	=1 if Member of a trade union	0.381	0.403	0.434
		(0.485)	(0.490)	(0.495)
health	=1 if bad health	0.125	0.075	0.095
		(0.330)	(0.264)	(0.294)
Number of Obs		937		

Figure X: Plot of Average Health against Average Income, by Individual



between the average health and average income of an individual. This suggests that health could be used in an earnings determination equation. However, for it to be an instrument, the orthogonality assumption implies that health should only determine income through education.

There is some doubt as to whether this orthogonality condition holds. It seems likely that poor health could lead to poor work performance (such as lower productivity or increased absenteeism) and thus lower earnings. However, in the coding of the health variable, three out of the five categories of health were coded as "good health", and only really poor health was coded as "bad health". Thus, it is likely that many of the illnesses suffered by those with "bad health" are serious and chronic problems, which could have affected their educational attainment, and may not have an additional affect on earnings outside of education. However, there is no way to be certain that there is no external affect of health on earnings. As such, the results must be viewed with some caution.

TABLE VII: RETURNS TO SCHOOLING, HAUSMAN-TAYLOR APPROACH

	Random Effects	Hausman Taylor
educ	0.070***	0.210***
	(0.006)	(0.019)
age	0.049***	0.075***
	(0.014)	(0.022)
agesq	-0.000*	-0.000
	(0.000)	(0.000)
female	-0.340***	-0.405***
	(0.042)	(0.055)
black	-0.620***	-0.310*
	(0.092)	(0.131)
coloured	-0.617***	-0.258
	(0.098)	(0.140)
asian	-0.204	-0.064
	(0.191)	(0.255)
rural	-0.147***	0.045
	(0.043)	(0.061)
union	0.240***	0.060
	(0.041)	(0.047)
health	-0.166***	-0.051
	(0.050)	(0.053)
Constant	6.520***	3.921***
	(0.327)	(0.587)
Observations	937	937

Marginal effects; Standard errors in parentheses

Source: NIDS

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3.2 Hausman-Taylor Estimation Results

Table VI presents the results of the HT estimation, with education treated as endogenous. The time-invariant exogenous variables are the race and gender dummies. The time-varying exogenous variables include age, rural, union and the industry and occupation dummies. A dummy variable for health is also included as a time-varying exogenous variable. This is in accordance with the original analysis done by Hausman and Taylor (1981) on returns to schooling. As such, under the HT approach, education is instrumented for by the individual means of the time-varying exogenous variables.

The estimated return to schooling has increased substantially with the HT method. It is now highly significant, with an additional year of schooling expected to increase wages by 21% on average for the working sample, when holding all else constant. A greater discussion into the factors driving this result will follow in the subsequent section. Other interesting differences between the random effects model and the HT estimator is the reduction in size of the racial wage differences, but the increase in the size of the gender wage difference. Furthermore, having bad health appears to have proportionately less effect on wages with the HT approach than random effects.

The union premium falls under the HT approach from 24% under random effects to 6%. However, the union variable is cause for concern because being in a union may be correlated with the omitted variables of ability and family background, and may thus be subject to endogeneity bias. As such, union is included as an endogenous variable together with education in Table VIII. As the results show, the union coefficient drops even further to 2.9%. This suggests that when accounting for observed worker characteristics, and unobserved background and ability differences, union membership does not have a significant wage premium in South Africa. Future study could look more closely at this union premium in South Africa under a panel data context.

In Table IX, education is measured in an alternative format, with dummy variables for educational attainment. The completion of Matric acts as the base group. Each of these dummies is treated as an endogenous time-invariant regressor. The HT specification suggests that having only primary or some secondary schooling has significantly lower returns than a Matric qualification. These coefficients are much larger than the corresponding random effect estimates. In turn, tertiary education yields a wage gain over a Matric qualification, but this effect is not as large. Again, differences in wages between racial groups appears to be less severe under HT, but the gender wage difference has not changed much between the two specifications.

TABLE VIII: RETURNS TO SCHOOLING, HAUSMAN-TAYLOR APPROACH WITH ENDOGENOUS UNION

	Random Effects	Hausman Taylor
educ	0.072***	0.191***
	(0.006)	(0.016)
age	0.047***	0.059**
	(0.014)	(0.023)
agesq	-0.000*	-0.000
	(0.000)	(0.000)
female	-0.328***	-0.402***
	(0.039)	(0.053)
black	-0.700***	-0.433***
	(0.088)	(0.128)
coloured	-0.722***	-0.419**
	(0.094)	(0.137)
asian	-0.326	-0.172
	(0.186)	(0.253)
rural	-0.135**	0.029
	(0.041)	(0.058)
union	0.221***	0.029
	(0.036)	(0.044)
health	-0.123**	-0.055
	(0.046)	(0.046)
Constant	6.684***	4.662***
	(0.318)	(0.576)
Observations	937	937

Marginal effects; Standard errors in parentheses

Source: NIDS

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE IX: HAUSMAN TAYLOR WITH EDUCATIONAL DUMMIES

	Random Effects	Hausman Taylor
primary	-0.570***	-1.251***
	(0.064)	(0.004)
secondary	-0.429***	-1.301***
	(0.056)	(0.005)
tertiary	0.432***	0.542***
	(0.072)	(0.005)
age	0.045***	0.079***
	(0.014)	(0.000)
agesq	-0.000*	-0.001***
	(0.000)	(0.000)
female	-0.337***	-0.356***
	(0.039)	(0.001)
black	-0.578***	-0.198***
	(0.088)	(0.002)
coloured	-0.580***	-0.171***
	(0.094)	(0.003)
asian	-0.323	-0.007*
	(0.183)	(0.003)
rural	-0.170***	-0.222***
	(0.040)	(0.001)
union	0.222***	0.018***
	(0.036)	(0.001)
health	-0.119**	0.009***
	(0.046)	(0.001)
Constant	7.540***	6.802***
	(0.301)	(0.010)
Observations	937	937

Marginal effects; Standard errors in parentheses

Source: NIDS

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3.3 Interpretation of the HT Results

Overall, the return to schooling for the main specification in Table VII is high at 21%. Possible explanations behind this high return must thus be given. Several factors that are driving this result will now be discussed in more depth. Firstly, as has been mentioned, this result is highly sample-specific. The sample includes those who consistently have work for a five-year period. However, this is likely to be a fairly elite group, as South Africa is characterised by high unemployment. According to Altman and Marock (2008), 58% of labour force participants aged 15-19 and 50% of individuals aged 20-24 are unemployed. As such, there may be characteristics, such as reliability, that determine the entry into the labour market and then the acquisition of stable employment, which are not adequately controlled for in the analysis. These characteristics may imply that higher education levels are more highly remunerated relative to those with lower education of this subgroup, relative to the overall population.

A second factor to be considered in these results is the lack of a control for the quality of schooling. As was mentioned earlier, the South African school system is polarized into institutions that were traditionally for blacks and others that were for whites. The lower quality of education in the former is an enduring legacy of this system. It may be the case that workers who are in stable employment are more likely to come from higher quality educational institutions. As such, these institutions may impart characteristics that get rewarded in the form of greater job stability and retention. It must be noted that the quality of the schooling institution is only likely to play a role in the analysis in terms of stable job acquisition and thus entry into the narrow balanced sample. If, however, quality of school affects the individual's time-constant characteristics, this has already been controlled for with the HT approach.

Thirdly, this sample of continuously employed individuals is likely to have had greater on-the-job training and skills development. It may be that highly educated individuals are given greater opportunities for job training and skills development. As such, employers may be hesitant to lose stable long-term workers and may therefore compensate them more highly. They may be rewarded more highly for this job-specific expertise relative to those with lower education levels. However, this effect would only materialise if within this sample of continuously employed individuals, the more highly educated were remunerated more relative to the lower educated of this continuously employed sample, compared to the wider population. On-the-job training will be considered in more detail in the following section.

A forth possible driver behind these high returns is the developing country context of South

Africa. Studies have consistently shown that developing countries reward education more highly than developed countries. Psacharopoulos and Patrinos (2004) find that in Sub-Saharan African countries, the return to primary education is 37%, and for secondary education it is 24%. This is significantly higher than the 13% and 11% for OECD countries respectively. Furthermore, as mentioned previously, Montenegro and Patrinos (2014) find that South Africa has the second highest return to education out of 139 countries. The pattern of higher returns for lower income countries is likely due to a shortage of skills in key areas of technology application and other professional service industries.

This finding of high returns to schooling is in line with the other aforementioned studies which similarly saw an increase in the estimated returns when using an instrumental HT approach. Similarly, it accords with the general instrumental literature of augmented returns to schooling when using IVs. It appears that the random effects estimate is biased downward through a lack of control for the unobserved individual effects. Ability and family bias exerted a seemingly negative effect on returns to schooling - more able individuals choose less schooling. Similarly, individuals with well-connected families or supportive family environments may feel less incentivized to gain higher education. More able or well-connected individuals may thus leave school earlier as they can more easily reach their target wage.

A possible explanation behind the increase in estimates is the hypothesis of sub-groups that was set out by Card (1999). As explained earlier, Card (1999) starts with the assumption that individuals with the lowest schooling tend to have the highest marginal return to schooling. If individuals in this subgroup dominate the instrument, then the return to schooling could rise. In the HT case, the instruments used are the individual means of age, rural, union, health and the industry and occupation dummies. It is not clear if any particular subgroup is dominating in this case. As such, this intuition is not applicable for the HT approach.

In summary, the HT approach yielded a high estimate of the return to schooling. This is likely due to the specific sample of employed workers across all three waves, who may have particular characteristics or skills that are highly rewarded with job stability. The context of a shortage of skills in South Africa may also be driving unequal wages in the labour market, and consequently the high returns for an additional year of education. The HT return is significantly higher than the random effects estimate, which may suggest that the omitted variables are biasing the random effects estimate downwards.

5 The Validity of the Results

5.1 On-the-job Training

This analysis assumed that the omitted individual heterogeneity is time-constant and therefore can be differenced out over time. This assumption seems plausible for individual characteristics such as intelligence, overall motivation and ambitiousness that are likely to be constant over time. Even if they do shift gradually, they are likely to be fairly stable for the five-year period that this analysis is limited to. However, there may be some features of individual heterogeneity that change over time. The key example is when an individual receives on-the-job-training, and therefore has large productivity increases over the years of the dataset. This omitted feature of ability is not eliminated with differencing and is therefore present in the error term. However, it is only a problem in the analysis if these productivity increases are correlated with education, which may not necessarily be the case.

A further way to check for productivity increases is to include a dummy for if the individual has remained in the same industry and occupation over the three waves. This therefore acts as a rough control for ability increases from job-training. Table X presents the analysis with these controls for job training. The results do not differ significantly from the previous coefficients. Returns to education decrease slightly to 19%, suggesting that productivity increases may have an effect on wages outside of the time-constant omitted variables. However, it appears that being in the same occupation has a negative effect on earnings, whilst being in the same industry has a wage gain. These confusing results may be partly due to the rather vague categories of industries and occupations in NIDS, where quite diverse jobs and industries are grouped together. As such, it is difficult to make any definitive conclusions from this.

5.2 Attrition

Attrition is typically a problem in a panel data context. In this study, the sample was narrowed down to a small group of workers who had permanent or casual employment over all three waves. As such, this sample neglects those individuals whose employment status changed over the three waves, which is a form of attrition. Additionally, the NIDS second wave had greater overall levels of attrition than the third wave because a smaller proportion of the original sample were contacted and re-surveyed in the second wave (NIDS Wave 3

TABLE X: RETURNS TO SCHOOLING, HAUSMAN-TAYLOR APPROACH WITH DUMMIES FOR SAME INDUSTRY AND OCCUPATION

	Random Effects	Hausman Taylor
educ	0.071***	0.192***
	(0.006)	(0.016)
age	0.048***	0.071**
	(0.014)	(0.024)
agesq	-0.000*	-0.000
	(0.000)	(0.000)
female	-0.331***	-0.396***
	(0.039)	(0.062)
black	-0.702***	-0.374**
	(0.088)	(0.143)
coloured	-0.723***	-0.399**
	(0.093)	(0.150)
asian	-0.329	-0.197
	(0.189)	(0.288)
rural	-0.135***	0.023
	(0.041)	(0.063)
union	0.224***	0.019
	(0.036)	(0.043)
health	-0.124**	-0.049
	(0.046)	(0.046)
sameoccup	0.034	-0.395**
	(0.040)	(0.136)
sameindustry	-0.035	0.373*
	(0.039)	(0.155)
Constant	6.696***	4.307***
	(0.317)	(0.600)
Observations	937	937

Marginal effects; Standard errors in parentheses

Source: NIDS

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Overview, 2013). As such, solely considering wave 1 and wave 3 will allow gains in sample size as it reduces these problems of attrition. Additionally, as pointed out by Hausman and Taylor (1981), limiting the analysis to two years minimises potential problems of serial correlation over the waves.

The sample size when using the first and third waves is 1364 individuals, which is a 45% increase in the sample size. This is after removing individuals from the sample who reported changing education levels. With this increased sample size, the returns to education appear to have increased somewhat, as the results in Table XI show. An additional year of education is expected to yield a 22% increase in education on average, holding all else constant. This is up from 21% for the analysis over all three waves.

TABLE XI: RETURNS TO SCHOOLING, HAUSMAN-TAYLOR APPROACH USING WAVE 1 AND 3

	Random Effects	Hausman Taylor
educ	0.076***	0.224***
	(0.006)	(0.016)
age	0.054***	0.084***
	(0.011)	(0.017)
agesq	-0.000**	-0.001**
	(0.000)	(0.000)
female	-0.328***	-0.391***
	(0.035)	(0.048)
black	-0.639***	-0.269*
	(0.072)	(0.108)
coloured	-0.619***	-0.203
	(0.077)	(0.115)
asian	-0.151	0.102
	(0.153)	(0.214)
rural	-0.154***	-0.028
	(0.037)	(0.097)
union	0.218***	-0.013
	(0.034)	(0.045)
health	-0.151***	-0.029
	(0.042)	(0.044)
Constant	6.427***	3.584***
	(0.258)	(0.471)
Observations	1364	1364

Marginal effects; Standard errors in parentheses

Source: NIDS

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.3 Validity of Instrument Choice

Hausman and Taylor (1981) suggested a specification test of the validity of their method. Under the null, the coefficients from the HT method are consistent. Similarly, the fixed-effects method is consistent because all variables are treated as endogenous and their correlation with the individual effect is eliminated. However, if HT is the more appropriate method, then fixed effects estimation is not as efficient as HT (Baltagi et al., 2003). It is these assumptions that are used as the basis for the HT test. This null hypothesis can also be represented as follows,

$$\underset{N\to\infty}{plim}\frac{1}{N}X_i'c = 0 \; and \; \underset{N\to\infty}{plim}\frac{1}{N}Z_i'c = 0 \qquad (15)$$

which implies that if the HT method is consistent, then all regressors are orthogonal to the individual effect. Under the alternative hypothesis, at least one of these orthogonality conditions does not hold. The test statistic derived by Hausman and Taylor (1981) is:

$$[\hat{\beta}_{HT} - \hat{\beta}_{FE}]'[Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{HT})]^{-1}[\hat{\beta}_{HT} - \hat{\beta}_{FE}] \xrightarrow{d} \chi^2(df) \qquad (16)$$

where the degrees of freedom are the number of time-varying exogenous variables minus the number of time-invariant endogenous regressors.

This test is applied to the main specification in Table VII. It yields a p-value of 0.9. The null hypothesis cannot be rejected, and so the Hausman-Taylor approach appears to be valid. However, the difficulty with this test is that the HT estimates are only compared against the estimates of the time-variant fixed effects regressors. A fixed effects approach does not generate estimates on education or race variables because they are time-constant. As such, this test is unable to act as a full comparison of the two models, and it fails to compare the coefficients on the variable of interest, namely education.

Furthermore, this test is not particularly robust, and is highly sensitive to the choice of instruments. In the HT approach, the partition of regressors into exogenous and endogenous variables is up to the researcher. This necessitates an assumption that the time-varying exogenous variables are uncorrelated with the individual effect. However, such an assumption is quite difficult to make. Even in the original Hausman and Taylor (1981) paper, there is doubt about their use of union and experience as exogenous variables, as these variables are likely to be correlated with individual characteristics such as motivation and ambitiousness. Wooldridge (2002: 361) thus comments on their paper, "It is difficult to know what to conclude, as the identifying assumptions are not especially convincing".

The difficulty in choosing exogenous instruments has been the source of a debate between Jordhal, Poutvaara and Tuomala (2009) and Garcia and Montuenga (2009). The former argue that the results are highly sensitive to the partition of regressors, which then means that the Hausman test is unstable and doesn't allow any convincing inferences. According to their analysis, when they used marriage as a variable, the test concluded that it was an exogenous variable. However, when a divorce dummy was used instead, the test determined that it was endogenous. Furthermore, these authors question whether variables that are typically assumed to be exogenous, such as occupation, are indeed uncorrelated with the unobserved effect. They argue that the instrument choice can lead to some of the strange results that Garcia and Monetuenga (2005) obtained, such as negative returns to a university education for self-employed individuals relative to the returns for secondary education.

The sensitivity of the HT method to the partition of regressors is similarly seen in this paper. In Table VII, union has a negative coefficient, which is unexpected given economic theory. As such, is is possible that this variable is correlated with the unobserved term and is being biased downwards. However, when adding union to the time-varying endogenous regressors in Table VIII, the Hausman-Taylor test yields a p-value of 0. This suggests that the Hausman specification is no longer valid which is unexpected. A similar occurrence happens when occupation is treated as endogenous. As such, it is difficult to make any strong conclusions from the Hausman-Taylor test.

A further difficulty is that some of the variables that are considered to be time-varying may actually show limited variation and thereby lead to the strange results. Garcia et al. (2009) argue that is a problem behind Jordhal's married and divorced results - these variables do not have enough variation over the short panel to adequately identify them. This may similarly be driving some of the strange results in this paper, as health, occupation, and the rural dummy may not have sufficient variation to capture the exogenous variation in education.

Overall, this discussion has highlighted the need to apply caution in reading the Hausman-Taylor results. A few steps can guide future studies that use a Hausman-Taylor approach. Firstly, the researcher needs to be sure that his partition of variables into exogenous and endogenous categories is theoretically sound. Secondly, the time-varying exogenous variables need to contain substantial variation to generate valid coefficients. Thirdly, the Hausman-Taylor test can shed some light on the validity of the model, but it fails to be the panacea in determining the partition of variables into endogenous and exogenous categories.

6 Conclusion

This paper has considered returns to schooling in South Africa. This is an important question in the South African context, where the government spends more than most other developing countries on education (Glewwe et al., 2011). The results show that this investment does appear to yield wage gains, as returns to schooling are approximately 21%. In particular, a Matric certificate seems to generate the most significant wage benefits compared to other levels of education.

The results in this paper were derived from using the Hausman-Taylor (1981) methodology. This approach has not been applied to South African data on education before. This paper therefore represents one of the first studies of returns to education using a national panel dataset in South Africa. The main benefit of the HT approach is its use of instruments internal to the dataset that are derived through time-differencing. This deals with the both the endogeneity of the education and the fact that education contains little time variation for the labour force sample.

The HT result of 21% is much higher than the cross-sectional OLS estimate of 7%. This is a common occurrence in returns to schooling studies that use an instrumental approach. It suggests the possibility that the omitted individual effect was causing a downward bias on the education coefficient. Controlling for this endogeneity leads to a rise in estimated returns to education. This implies that ability and family background tend to offset the choice in education levels. It may thus be that family connections and natural ability play a strong role in South Africa, and individuals do not rely as much on a high education level to obtain target wages.

A second reason for the high return to education under a HT approach is that the panel sample in this study was narrowed down to individuals with permanent or casual employment over the entire 5-year period in which the survey took place. This implies that there is a strong selection effect in action. It may be that characteristics which are keeping the individual in the job, such as job-specific skills, are more highly rewarded for the educated relative to the lower-educated in the sample, compared to this difference in the wider population sample.

Overall, the HT results must be viewed with some caveats. It is difficult to motivate for time-varying instruments that are completely orthogonal to the individual's unobserved characteristics. Additionally, these variables need to have substantial variation, which may be challenging to obtain. However, with these caveats in mind, the results suggest returns to

education are very high for the permanently employed. Permanent employment and job stability is therefore a channel by which inequality in South Africa is reinforced. Not only are the permanently employed receiving a constant income, but the increase in income for each education level they obtain is very high. This is a double whammy for individuals who have both periods of unemployment and little access to high levels of education.

7 Bibliography

Allison, P. D. (2001). Missing data. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Altman, M., & Marock, C. (2008). Identifying appropriate interventions to support the transition from schooling to the workplace. Human Sciences Research Council Overview Discussion Paper, 1.

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

Arcand, J., d'Hombres, B., & Gyselinck, P. (2004). Instrument choice and the returns to education: New evidence from Vietnam. Econ WPA Working Paper, No. 200422.

Arias, O., Hallock, K. F., & Sosa-Escudero, W. (2002). Individual heterogeneity in the returns to schooling: instrumental variables quantile regression using twins data. In Economic Applications of Quantile Regression (pp. 7-40). Physica-Verlag HD.

Baltagi, B. H., Bresson, G., & Pirotte, A. (2003). Fixed effects, random effects or Hausman–Taylor?: A pretest estimator. Economics letters, 79(3), 361-369.

Baltagi, B. H., & Bresson, G. (2012). A robust Hausman–Taylor estimator. Advances in Econometrics, 29, 175-214.

Bandiera, O., & Rasul, I. (2006). Social networks and technology adoption in northern Mozambique. The Economic Journal, 116(514), 869-902.

Banerjee, A., Galiani, S., Levinsohn, J., McLaren, Z., & Woolard, I. (2008). Why has unemployment risen in the new South Africa? Economics of Transition, 16(4), 715-740.

Battistin, E., De Nadai, M., & Sianesi, B. (2012). Misreported schooling, multiple measures and returns to educational qualifications. Discussion Paper Series, Forschungsinstitut zur Zukunuft der Arbeit, No. 6337.

Becker, G. (1964). Human capital: A theoretical and empirical analysis, with special reference to education. Columbia University Press: New York.

Behrman, J. R., & Wolfe, B. L. (1984). The socioeconomic impact of schooling in a developing country. The Review of Economics and Statistics, 296-303.

Belzil, C., & Hansen, J. (2002). Unobserved ability and the return to schooling. Econometrica, 70(5), 2075-2091.

Bilbao-Osorio, B., Dutta,S., & Lanvin, B., Editors (2014). The global information technology report, 2014. World Economic Forum insight report.

Besley, T., & Case, A. (1993). Modeling technology adoption in developing countries. The American Economic Review, 396-402.

Bhorat, H. (2000). Wage premia and wage differentials in the South African labour market. DPRU Working Papers, 43.

Branson, N., Kekana, D., & Lam, D. (2013). Educational expenditure in South Africa: Evidence from the National Income Dynamics Study. South African Labour and Development Research Unit Working Paper 124.

Bound, J., & Jaeger, D. A. (1996). On the validity of season of birth as an instrument in wage equations: A comment on Angrist and Krueger's 'Does compulsory school attendance affect schooling and earnings?'. NBER Working Paper #5835.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. Handbook of econometrics, 5, 3705-3843.

Bowles, S., H. Gintis and M. Osborne. (2001). The determinants of earnings: a behavioural approach. Journal of Economic Literature 39, 1137–76.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling (No. w4483). National Bureau of Economic Research.

Card, D. (1996). The effect of unions on the structure of wages: A Longitudinal Analysis. Econometrica 64(4), 957–979.

Card, D. (1999). The causal effect of education on earnings. Handbook of labor economics, 3, 1801-1863.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. Econometrica, 69(5), 1127-1160.

Casale, D. & Posel, D. (2005). Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa. Mimeo, Durban: University of Kwazulu-Natal

Case, A., & Yogo, M. (1999). Does school quality matter? Returns to education and the characteristics of schools in South Africa. National Bureau of Economic Research, Working Paper, No. w7399.

Centre for Development and Enterprise. (2013). Affordable private schools in South Africa. August 2013 report.

Chamberlain, D., & Van der Berg, S. (2002). Earnings functions, labour market discrimination and quality of education in South Africa. Stellenbosch University, Department of Economics, Working Paper, No. 02.

Daniels, R. (2008). The income distribution with coarse data. Economic Research Southern Africa, Working Paper, (82).

Daniels, R. (2012). Univariate Multiple Imputation for Coarse Employee Income Data. South African Labour and Development Research Unit, Working Paper, No. 88. Cape Town: University of Cape Town.

De Villiers, L., Brown, M., Woolard, I., Daniels, R.C., & Leibbrandt, M, eds. (2013). National Income Dynamics Study Wave 3 User Manual. Cape Town: Southern Africa Labour and Development Research Unit.

Dupas, P. (2014). Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence From a Field Experiment. Econometrica, 82(1), 197-228.

Falaris, E. M. (1995). The role of selectivity bias in estimates of the rate of return to schooling: the case of married women in Venezuela. Economic Development and Cultural Change, 43(2), 333-350.

Fryer, D., & Vencatachellum, D. (2005). Returns to education in South Africa: evidence from the Machibisa township. African Development Review, 17(3), 513-535.

Garcia-Mainar, I., & Montuenga-Gomez, V. M. (2005). Education returns of wage earners and self-employed workers: Portugal vs. Spain. Economics of Education Review, 24(2), 161-170.

Garcia-Mainar, I., & Montuenga-Gomez, V. M. (2009). Education returns of wage earners and self-employed workers: Response. Economics of Education Review, 28(5), 645-647.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010 (No. w17554). National Bureau of Economic Research.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. Econometrica: Journal of the Econometric Society, 1-22.

Gustafsson, M., & Mabogoane, T. (2012). South Africa's economics of education: A stocktaking and an agenda for the way forward. Development Southern Africa, 29(3), 351-364.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. Journal of economic literature, 607-668.

Harmon, C., & Walker, I. (1995). Estimates of the economic return to schooling for the United Kingdom. The American Economic Review, 85(5), 1278-1286.

Harmon, C., Oosterbeek, H., & Walker, I. (2000). The returns to education: a review of evidence, issues and deficiencies in the literature. Centre for the Economics of Education, London School of Economics and Political Science.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. Econometrica: Journal of the Econometric Society, 1377-1398.

Heckman, J. J., & Hotz, V. J. (1986). An investigation of the labor market earnings of panamanian males evaluating the sources of inequality. Journal of Human Resources, 507-542.

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. Handbook of the Economics of Education, 1, 307-458.

Heitjan, Daniel F. & Donald B. Rubin. (1991). Ignorability and coarse Data. The Annals of Statistics, 1991, 19 (4), 2244—2253.

Hertz, T. (2003). Upward bias in the estimated returns to education: Evidence from South Africa. The American Economic Review, 93(4), 1354-1368.

Hungerford, T. & Solon, G. (1987). Sheepskin effects in the return to education. Review of Economics and Statistics, 69, 175-177.

Johansson, F. (2005). A measurement error analysis of survey data–Using administrative data as a validation source. Uppsala Univsersity Working Paper.

Jones, P. (2001). Are educated workers really more productive?. Journal of Development Economics, 64(1), 57-79.

Jordahl, H., Poutvaara, P., & Tuomala, J. (2009). Education returns of wage earners and self-employed workers: Comment. Economics of Education Review, 28(5), 641-644.

Kane, T. J., Rouse, C. E., & Staiger, D. (1999). Estimating returns to schooling when schooling is misreported (No. w7235). National Bureau of Economic Research.

Kermyt G. A., Case, A. & Lam, D. (2001) Causes and consequences of schooling outcomes in South Africa: Evidence from survey data. Social Dynamics: A

journal of African studies, 27:1, 37-59.

Keswell, M. (2004). Education and racial inequality in post Apartheid South Africa. Sante Fe Institute, Working Paper, No. 02-008.

Keswell, M., & Poswell, L. (2004). Returns to education in South Africa: A retrospective sensitivity analysis of the available evidence. South African Journal of Economics, 72(4), 834-860.

Lacerda, M., Ardington, C., & Leibbrandt, M. (2007). Sequential regression multiple imputation for incomplete multivariate data using Markov Chain Monte Carlo.

Lam, D., & Schoeni, R. F. (1993). Effects of family background on earnings and returns to schooling: evidence from Brazil. Journal of political economy, 710-740.

Mariotti, M., & Meinecke, J. (2014). Partial identification and bound estimation of the average treatment effect of education on earnings for South Africa. Oxford Bulletin of Economics and Statistics, 0305-9409.

McCord, A., & Bhorat, H. (2003). Employment and labour market trends. Human resources development review, 112-41.

Millimet, D. L. (2011). The elephant in the corner: a cautionary tale about measurement error in treatment effects models (Vol. 27, pp. 1-39). Emerald Group Publishing Limited.

Mincer, J. A. (1974). Schooling and earnings. In Schooling, experience, and earnings (pp. 41-63). Columbia University Press.

Montenegro, C. E., & Patrinos, H.A. (2014). Comparable estimates of returns to schooling around the world. World Bank policy research working paper 7020.

Oreopoulos, P., & Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. The Journal of Economic Perspectives, 159-184.

Psacharopoulos, G., & Patrinos, H. A. (2004). Returns to investment in education: a further update. Education economics, 12(2), 111-134.

Rosenzweig, M. R. (1995). Why are there returns to schooling? American Economic Review, 85(2), 153-158.

Rosenzweig, M. R. (2010). Microeconomic approaches to development: Schooling, learning, and growth. The Journal of Economic Perspectives, 81-96.

Rubin, D. (1976). Inference and missing Data. Biometrica, 63, 581- 592.

Rubin, D. (1987) Multiple imputation for nonresponse in surveys. New York: Wiley.

Serumaga-Zake, P. A., & Naude, W. A. (2003). Private rates of return to education of Africans in South Africa for 1995: a Double Hurdle model. Development Southern Africa, 20(4), 515-528.

Southern Africa Labour and Development Research Unit. (2013). National Income Dynamics Study, Wave 3,2,1 [dataset]. Cape Town: Southern Africa Labour and Development Research Unit [producer]. Cape Town: DataFirst [distributor].

Southern Africa Labour and Development Research Unit. (2013). National Income Dynamics Study Wave 3 Overview. Cape Town: Southern Africa Labour and Development Research Unit [producer]. Cape Town: DataFirst [distributor].

Spaull, N., & Taylor, S. (2012). Effective enrolment–Creating a composite measure of educational access and educational quality to accurately describe education system performance in sub-Saharan Africa. Stellenbosch Economic Working Paper (No. 21/2012).

Spaull, N. (2013). South Africa's Education Crisis: The quality of education in South Africa 1994-2011. Centre for Development and Enterprise Report.

Swaffield, J. K. (2001). Does measurement error bias fixed-effects estimates of the union wage effect?. Oxford Bulletin of Economics and Statistics, 63(4), 437-457.

Vermaak, C. (2010). The impact of multiple imputation of coarsened data on estimates on the working poor in South Africa (No. 2010, 86). Working paper//World Institute for Development Economics Research.

Wittenberg, M. (2008). Nonparametric estimation when income is reported in bands and at points. Cape Town: Economic Research Southern Africa Working Paper, (94).

Wittenberg, M. (2013). A comment on the use of "cluster" corrections in the context of panel data. National Income Dynamics Study Note, August.

Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data. Cambridge: Massachusetts Institute of Technology Press.

Wooldridge, J. M. (2012). Introductory econometrics: A modern approach, 5th edition. South Western: Cengage Learning.

Wright, R. E. (1999). The rate of return to private schooling (No. 92). IZA Discussion paper series.