Labour Market Returns to Educational Attainment, School Quality, and Numeracy in South Africa

by

Hendrik van Broekhuizen

Thesis presented in partial fulfilment of the requirements for the degree of Master's of Commerce at the University of Stellenbosch

Supervisor: Prof. Servaas van der Berg
Faculty of Economic and Management Sciences
Department of Economics

December 2011

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

December 2011

Summary

This study investigates the extent to which educational attainment, school quality and numeric competency influence individuals' employment and earnings prospects in the South African labour market using data from the 2008 National Income Dynamics Study (NIDS). While NIDS is one of the first datasets to contain concurrent information on individual labour market outcomes, educational attainment levels, numeric proficiency and the quality of schooling received in South Africa, it is also characterised by limited and selective response patterns on its school quality and numeracy measures. To account for any estimation biases that arise from the selective observation of these variables or from endogenous selection into labour force participation and employment, the labour market returns to human capital are estimated using the Heckman Maximum Likelihood (ML) approach. The Heckman ML estimates are then compared to Ordinary Least Squares (OLS) estimates obtained using various sub-samples and model specifications in order to distinguish between the effects that model specification, estimation sample, and estimation procedure have on estimates of the labour market returns to human capital in South Africa.

The findings from the multivariate analysis suggest that labour market returns to educational attainment in South Africa are largely negligible prior to tertiary levels of attainment and that racial differentials in school quality may explain a significant component of the observed racial differentials in South African labour market earnings. Neither numeracy nor school quality appears to influence labour market outcomes or the convex structure of the labour market returns to educational attainment in South Africa significantly once sociodemographic factors and other human capital endowment differentials have been taken into account. Though the regression results vary substantially across model specifications and estimation samples, they are largely unaffected by attempts to correct for instances of endogenous selection using the Heckman ML procedure. These findings suggest that the scope for overcoming data deficiencies by using standard parametric estimation techniques may be limited when the extent of those deficiencies are severe and that some form of sensitivity analysis is warranted whenever data imperfections threaten to undermine the robustness of one's results.

Opsomming

Hierdie studie ondersoek in watter mate opvoedingspeil, skoolgehalte en numeriese vaardighede individue se werks- en verdienstevooruitsigte in die Suid-Afrikaanse arbeidsmark beïnvloed. Die studie gebruik data van die 2008 *National Income Dynamics Study* (NIDS). Alhoewel NIDS een van die eerste datastelle is wat inligting oor individuele arbeidsmarkuitkomste, opvoedingsvlakke, numeriese vaardighede sowel as skoolgehalte bevat, word dit ook gekenmerk deur beperkte en selektiewe responspatrone rakende skoolgehalte en die numeriese vaardigheidmaatstaf. Die arbeidsmarkopbrengs op menslike kapitaal word deur middel van die Heckman 'Maximum Likelihood (ML)'-metode geskat om te kontroleer vir moontlike sydighede wat mag onstaan weens selektiewe waarneming van hierdie veranderlikes of as gevolg van endogene seleksie in arbeidsmarkdeelname of indiensneming. Die Heckman ML-skattings word dan vergelyk met gewone kleinste-kwadrate-skattings wat met behulp van verskeie modelspesifikasies en steekproewe beraam is, om sodoende te bepaal hoe verskillende spesifikasies, steekproewe en beramingstegnieke skattings van die arbeidsmarkopbrengste op menslike kapitaal in Suid-Afrika beïnvloed.

Die meerveranderlike-analise dui daarop dat daar grotendeels onbeduidende arbeidsmarkopbrengste is op opvoeding in Suid-Afrika vir opvoedingsvlakke benede tersiêre vlak, en dat rasseverskille in skoolgehalte 'n beduidende deel van waargenome rasseverskille in arbeidsmarkverdienste mag verduidelik. Indien sosio-demografiese faktore en ander menslike kapitaalverskille in ag geneem word, beïnvloed syfervaardigheid en skoolgehalte nie arbeidsmarkuitkomstes en die konvekse struktuur van die arbeidsmarkopbrengste op opvoeding in Suid-Afrika beduidend verder nie. Terwyl die regressieresultate aansienlik tussen die verskillende modelspesifikasies en steekproewe verskil, word die resultate weinig geraak deur vir gevalle van endogene seleksie met behulp van die Heckman ML-metode te kontroleer. Hierdie bevindinge dui daarop dat daar net beperkte ruimte bestaan om ernstige dataleemtes met behulp van standaard parametriese beramingstegnieke te oorkom, en dat die een of ander vorm van sensitiwiteitsanalise benodig word wanneer datagebreke die betroubaarheid van die beraamde resultate nadelig kan raak.

Acknowledgements

The author would like to thank Prof. Servaas van der Berg for his financial assistance, academic support and continued guidance as well as Dieter von Fintel, Gideon du Rand and other members of the Social Policy Research Group at the Department of Economics at Stellenbosch University for their invaluable comments and suggestions. In places, this study sources extensively from Du Rand *et al.* (2010, 2011). Where formulations are those of co-authors Gideon du Rand and/or Dieter von Fintel the appropriate acknowledgement is given. Any errors remain the sole responsibility of the author.

Dedications

I dedicate this thesis to my best friend and wonderful girlfriend, Jenny Schnepper, and to my loving parents, Gawie and Susan van Broekhuizen. This work would not have been possible without your constant love, understanding, support, prayer, and many words of encouragement. I love you all.

Contents

List of Figures

List of Tables

List of Abbreviations

AMSQE	Average Marginal School Quality Effect
2SLS	Two-Stage Least Squares
DGP	Data Generating Process
HCT	Human Capital Theory
IMR	Inverse Mills Ratio
IV	Instrumental Variable(s)
LF/LFP	Labour Force/Labour Force Participation
MCA	Multiple Correspondence Analysis
MAR\NMAR	Missing-at-Random\Not-Missing-at-Random
ML/MLE	Maximum Likelihood/Maximum Likelihood Estimation
NIDS	National Income Dynamics Study
OLS	Ordinary Least Squares
ROC	Receiver Operating Curve
ROR	Rate(s) of Return
NSCE	National Senior Certificate Examinations
SH	Sorting Hypothesis

Chapter 1

Introduction

Nearly two decades after the transition to democracy, South Africa's labour market remains characterised by widespread inequality, persistently high unemployment and substantial variation in the labour market prospects faced by its working-age population. While the factors that have contributed and continue to contribute to this labour market landscape are complex and multifaceted, there is an increasing need for research that attempts to identify those areas where policy interventions are not only most crucial, but also stand to be most effective. Given the incontrovertible evidence regarding the substantial private labour market benefits accruing from investments in human capital in both the international and local literatures, such research invariably necessitates an investigation of the state and distribution of human capital within South Africa's labour force, the nature, quantity, quality and composition of individual human capital endowments, and the roles that specific indicators of human capital play separately and collectively in determining labour market outcomes in the country.

Using data from the 2008 National Income Dynamics Study (NIDS), one of the first nationally representative datasets that allows for various human capital indicators to be linked to individual labour market outcomes, this study contributes to the literature on the nature of the relationships between human capital investments and labour market outcomes in South Africa by examining the relative impacts of educational attainment, school quality and numeracy on the probability of being employed and expected earnings capacity in the South African labour market. The complex underlying relationships between the human capital and labour market outcome variables considered and the significant extent of selective non-observability on the NIDS school quality and numeracy measures suggest that it is necessary to correct for omitted variable and sample selection bias when estimating the employment and earnings returns to educational attainment, school quality and numeracy and imply that the estimation results are unlikely to be robust to different model specifications and estimation samples. In order to assess the robustness of the findings, the primary objective of the analysis is thus to produce not only one set of

point estimates, but a range of estimates of the employment and earnings returns to educational attainment, school quality and numeracy in the South African labour market. Several auxiliary hypotheses are also investigated. These relate to the extent to which controlling for school quality and numeracy in labour market returns estimations influence the convexity in the structure of the estimated returns to educational attainment and the magnitudes of the unexplained components in racially-delineated employment and earnings outcome differentials.

Based on the preliminary findings from the descriptive analysis, the aforementioned objectives are pursued by structuring the multivariate analysis within a bottom-up methodological framework in an attempt to isolate the effects that different model specifications, different estimation samples, and different estimation procedures have on the regression estimates. While the empirical results are subject to various caveats and deviate from *a priori* expectations and the findings of other studies in a number of important respects, they nevertheless provide some insight into the potential magnitudes of the private employment and earnings returns to school quality, numeracy, and educational attainment in South Africa. As such, the conclusions that can be drawn from the findings should be of value to both policy makers and researchers.

The results show that the South African labour market returns to education are negligible before tertiary levels of attainment, but are large and increasing thereafter. There is also circumstantial evidence to suggest that racial differentials in school quality may explain a significant component of the observed racial differentials in labour market earnings, thus supporting the notion that part of the labour market inequalities that are often attributed to persistent labour market discrimination may be rooted in pre-labour market inequalities in the South African education system. These findings imply that there is a need for educational policy to extend beyond the provision of access to education and focus on improving the quality of education, particularly in historically-Black schools.

In contrast to the findings from other studies, the results from the multivariate analysis suggest that numeracy and school quality do not significantly influence labour market outcomes or the structure of the labour market returns to educational attainment in South Africa once sociodemographic factors and other human capital endowment differentials have been taken into account. However, this result appears to be explained largely by the selective pattern of observations on the numeracy and school quality variables, the peculiar measurement of these variables, and the lack of precision in the estimations due to the small sizes of the samples within which the measures are respectively captured. The capacity for standard parametric sample selection correction procedures to compensate for these issues is shown to be limited given the severity of the deficiencies in the data. As such, the results most likely fail to accurately reflect the extent of the importance of numeracy and school quality for labour market outcomes in South Africa. Moreover, the sensitivity of the estimation results to different model specifications and estimation samples reveals the need for more accurate and representative data on human capital

endowments in South Africa and for greater transparency in the estimation, presentation, and interpretation of estimates of the labour market returns to imperfectly measured and selectively captured indicators of human capital.

To contextualise the discussion of the labour market returns to educational attainment, school quality and numeracy in South Africa, Chapter 2 defines the concepts of human capital and labour market returns to human capital and provides a conceptual overview of the underlying theoretical relationships between the different human capital components and labour market outcomes that are considered in the analysis. In addition, the existing literature and empirical evidence on the effects of educational attainment, school quality and numeracy on employment and earnings, both internationally and in the South African labour market, are reviewed.

Chapter 3 introduces the 2008 National Income Dynamics Study (NIDS) data which is used in the empirical analysis and discusses some of its most important features. While the NIDS data has the advantage of containing information on multiple human capital and labour market outcome indicators, it is revealed to have several potentially serious disadvantages. Foremost among these is the extremely limited number of non-missing observations on the NIDS numeracy and school quality variables, both of which are likely to be upward-biased indicators of actual numeracy and school quality levels in South Africa, given that they seem to inadequately capture the lower tails of the true school quality and numeracy distributions. Moreover, the patterns of observability on these two variables are shown to be systematically related to many of the observable determinants of labour market earnings and the probability of employment and there are indications that they may similarly be related to certain unobserved correlates of labour market outcomes as well.

The findings from Chapter 3 suggest that, in addition to normal concerns regarding omitted variable and sample selection bias that apply whenever labour market returns to educational attainment are estimated, it may also be necessary to explicitly control for any biases that arise from the selective observation of the NIDS numeracy and school quality variables in the empirical analysis. Following a brief review of the literature on the effects of and solutions to omitted variable bias and selection bias in the context of labour market returns estimations, Chapter 4 provides a formal description of these issues. In order to achieve the objectives of the empirical analysis, it is suggested that the NIDS school quality and numeracy score measures should be used to proxy for omitted variables in the estimation of the labour market returns to educational attainment and that the Heckman Maximum Likelihood (ML) sample selection correction procedure should be used to correct for the respective instances of potentially endogenous selection into the labour force participant, earnings, school quality and numeracy estimation samples.

As precursor to the multivariate analysis, Chapter 5 provides a descriptive overview of the states of, and relationships between, various sociodemographic factors, human capital endowments,

and labour market outcomes in South Africa. The findings from the descriptive analysis contextualise the estimation of the labour market returns to education, school quality, and numeracy in South Africa and provide a number of priors against which to evaluate the findings of the multivariate analysis. Among these are are the notions that school quality and numeracy not only have strong and positive associations with employment and earnings outcomes in South Africa, but also share strong and positive associations with educational attainment levels.

The methodological approach used in and the results of the multivariate analysis is presented in Chapter 6. Beginning with simple estimations and progressively adding complexity, the analysis is structured with the specific intent of disentangling the effects that different model specifications, estimation samples, and estimation procedures have on the estimates of the labour market returns to educational attainment, school quality and numeracy in South Africa. The results reveal that changes in the magnitudes and statistical significance of the various coefficient estimates are driven almost exclusively by systematic differences in the estimation samples and appear to be largely unaffected by any of the attempts to correct for endogenous sample selection using the Heckman ML procedure.

Lastly, Chapter 7 summarizes the main findings from the empirical analysis, discusses some important caveats pertaining to various theoretical considerations and practical issues that may undermine the validity and interpretability of those findings, and concludes on what the implications of the findings in this study are for the assessment of the labour market returns to different components of human capital in South Africa.

Chapter 2

Concepts, Theory and Existing Evidence: Human Capital and Labour Market Returns

The multi-dimensionality and abstract nature of the human capital concept implies that the theoretical linkages between investments in human capital and labour market outcomes are inherently complex. Given this complexity, it is generally difficult to disentangle the underlying causal relationships between human capital, labour market productivity, and labour market prospects without careful consideration of the existing theoretical and empirical literature. As precursor to the conceptual and empirical analyses presented in this study, the present chapter therefore commences with an overview of the key underlying theoretical considerations that govern the study of the labour market returns to human capital and summarises some of the existing evidence on the effects of educational attainment, school quality and numeracy on employment and earnings prospects, both internationally and in South Africa.

2.1 Human Capital

While the term *human capital* was first used by Arthur C. Pigou in 1928 in *A Study in Public Finance*, the notion that human capacity and faculty constitutes a form of capital long pre-dates the origin of this term (Pigou, 1928, p.29). In his 1776 opus *An inquiry into the Nature and Causes of the Wealth of Nations,* Adam Smith asserted that the chief components of society's stock of fixed capital included *"...the acquired and useful abilities of all the inhabitants or members of the society."* (Smith, 2009, p.166). Smith further described the origin, content and implications of this *human* component of capital:

> *"The acquisition of such talents, by the maintenance of the acquirer during his education, study, or apprenticeship, always costs a real expense, which is a capital fixed and realized, as it were, in his person. Those talents, as they make a part of his fortune, so do they likewise of that of the society to which he belongs. The improved dexterity of a workman may be considered in the same light as a machine or instrument of trade which facilitates and abridges labour, and which, though it costs a certain expense, repays that expense with a profit."*
> (Smith, 2009, p. 166)

Despite Smith's acknowledgement of human capital and his insightful description thereof more than two centuries before, the term *human capital,* along with the concepts it embodies and its use as theoretical justification for the observed relationship between education and labour market productivity only came to the forefront of the labour economics literature in the 1960's following the seminal works of Theodore W. Schultz and Gary S. Becker which were extended in the 1970's by Jacob Mincer, George Psacharopoulos and Mark Blaug (Becker, 1992, p.43).[1] In the subsequent decades, *human capital* proliferated as the subject of academic study and political interest, not only evolving in concept, but also becoming a generic conceptual description for the value of labour. It is therefore perhaps surprising that there is no single, encompassing and universally accepted definition of human capital. Yet, in order to understand its value in the labour market, some definition must be ventured.

At the aggregate level, the term *human capital* is sometimes used as a generic collective for the total potential productive capacity of all labour in a country, sector, industry, or firm. However, for the purpose of this study the focus falls on human capital as it operates at the level of the individual. All individuals possess a stock of human capital comprising of all the psychological and physiological experiences, attributes, and capacities that relate to the determination of their potential and realized labour market productivity and, consequently, the theoretical value of their labour. Human capital is therefore not only complex and somewhat abstract in nature, but also inherently difficult to measure.

From the definition provided here and those put forth by other authors, it is possible to identify four key aspects that characterise the nature of human capital. First, individuals' stocks of human capital are variable over time. The activities which people engage in, that which they observe, learn, study, and practice, and the way in which they adapt to circumstances invariably augment their existing stocks of human capital. So too is it possible for human capital to be destroyed because of injury and illness or as the natural consequence of the physical and mental decay associated with ageing. In other words, human capital is neither purely innate, nor simply static.

[1] See Schultz (1961, 1962, 1963), Becker (1962, 1964), Psacharopoulos (1973), Mincer (1974) and Blaug (1972, 1976).

Second, each individual's stock of human capital is a unique composition of its constituent components. Not only do some individuals have greater innate abilities and aptitudes than others, but the nature of those abilities and aptitudes also differ from one person to the next. However, differences in human capital stocks between individuals are not simply innate, but come as a direct consequence of the types of human capital that people choose to expand and acquire and the ways in which they choose to do so. An understanding of the human capital augmentation process is therefore important for understanding differences between individuals' human capital stocks. This has important practical implications as it is often easier to measure the value or the magnitude of the steps taken to augment human capital (e.g. the number of years of formal education completed or the number of books an individual has read) than it is to measure actual human capital itself.

Individuals possess both general and specific forms of human capital. General human capital, like literacy, is useful in a wide array of applications and allows for the performance of many types of labour. By contrast, specific human capital, like an advanced knowledge of Chinese maritime law or the ability to kick a football across the width of a football field, are only relevant to the performance of specific types of labour and have limited value in other contexts (Kerckhoff *et al.*, 2001, pp. 2-3). This highlights the third characterising aspect of human capital: the value of an individual's stock of human capital at any point in time is context-dependent. People are arguably more productive in occupations where their specific skills and competences are relevant to the tasks they perform. Therefore, the extent to which the types of human capital individuals possess are aligned with the nature of the labour they are expected to perform determines the worth of that human capital (Wolpin, 1977, p.950). The greater the compatibility, the smaller the divergence between their potential and actual labour market productivities and the greater the value of their human capital in that specific setting. The skills of a professional trapeze artists, for example, while certainly remarkable in their own right, are of little value to someone pursuing a career as a neurosurgeon. It follows that at any point in time an individual with a given stock of human capital may be highly productive in one job, and yet far less productive in another.

Finally, for any skill, aptitude, or characteristic to be defined as human capital, it must influence labour productivity. This raises the critical question of what precisely can be called human capital in practice, how it is acquired, and how it should be measured. Such questions are the source of considerable debate between social scientists and a vast number of studies have been dedicated to identifying feasible measures of human capital. To do so, it is necessary to shift the focus away from overly abstract conceptualisations of human capital and concentrate on common observables which should, in theory, be highly correlated with labour productivity.[2]

[2] Here, "observables" refer to factors which are generally easy to observe and comparatively easy to measure.

The most commonly studied indicators of human capital that are found in the literature can be divided into five broad categories: the scope, type, and quality of educational attainment; the nature and extent of labour market and labour market-related experiences; natural intelligence, capabilities, and other innate capacities; the extent and nature of specific acquired aptitudes and cognitive skills; and the nature and extent of emotional intelligence, motivation, and other non-cognitive skills. While the majority of studies focus on educational attainment as the foremost augmenter, reflector, predictor, and/or signal of human capital and attempt to draw causal links between educational attainment levels and labour market outcomes, there is an increasing tendency to include measure of education quality, measures of ability such as IQ or aptitude test scores, and measures of specific skills such as literacy or numeracy tests scores in empirical labour market analyses (Kingston *et al.*, 2003, p. 55). The present study continues this trend by focusing on several indicators of human capital as determinants of labour market outcomes in South Africa. Abstracting from the impact of labour market experience, the measures that are considered are the number of years of educational attainment, the quality of formal secondary schooling, and numeracy.[3]

2.2 Human Capital and Labour Market Returns

Given the explicit link between human capital and productivity, it is not difficult to appreciate that labour markets generally have greater demand for and more handsomely reward individuals who possess valuable human capital. Individuals' human capital stocks largely determine the labour market outcomes that they face and an expansion of human capital should, *ceteris paribus*, improve employment and remuneration prospects.[4] Specifically, the labour market benefits of human capital investments are expected to manifest in three major respects. First, it should increase the probability of procuring employment.[5] Second, it should increase the likelihood that the type of employment procured is compatible with the nature of an individual's specific human capital stock and provides greater on-the-job benefits and job security. Third, and following directly from the second point, the expansion of human capital should raise the expected earnings of individuals who are employed (Bhorat and McCord, 2003, p. 135).[6] Investments

[3] For the sake of simplicity, the empirical analyses in Chapters 5 and 6 abstract from modelling the costs associated with and the decisions underlying investments in human capital and instead assume that individuals' levels of educational attainment, quality of schooling received and numeracy are commensurate to the extent of their investments therein.

[4] In a world of asymmetric information and rigidities, human capital will not, of course, be the sole determinant of labour market outcomes. For example, some studies have found that social capital may be just as important for procuring employment as human capital (Knight and Yueh, 2002, p. 2).

[5] In this study, the employed includes formal, casual, private, public, and self-employed individuals.

[6] The realisation of these labour market benefits do not require the marginal productivity theory to hold. Even if labour were not paid its marginal product, there is sufficient evidence to suggest that, on average and with all else being held constant, more productive and more specialised labour is better remunerated than its less productive and more general counterpart (Blaug, 1976, p. 54).

in human capital thus generate certain private labour market returns for the individual.[7] This study specifically focusses on the increases in the probability of being employed and the subsequent earnings of individuals who are employed that can be ascribed to investments in the three components of human capital identified above. In the remainder of this paper these returns are respectively referred to as the *employment returns* and *earnings returns* to educational attainment, school quality, and numeracy.

2.2.1 Returns to Educational Attainment[8]

The positive association between education and labour market earnings is one of the most robust empirical findings in the economics of education and labour market literatures. While fewer studies have been devoted to assessment of the relationship between education and the probability of being employed, a similarly strong and robust positive association is commonly found to exist between the two outcomes (Bhorat and McCord, 2003, p. 135). Two primary theories have been advanced to explain the reason for these positive associations. The Human Capital Theory (HCT) asserts that education instils and expands such characteristics and capacities as fall within the ambit of human capital and thus implies a direct causal link between education and labour market productivity. In other words, investments in education yield labour market returns because education acts as an augmentation device by way of which innate abilities and aptitudes are moulded into such productive capacities as are valued in the labour market.

The Sorting Hypothesis (SH)[9] extends the HCT by postulating that education serves as a signal of critical information regarding individuals' innate productive capacities (Weiss, 1995, p. 134). Given that progression through education requires the successful completion of a series of competency-based tasks, part of its implied function is to sort individuals according to their abilities to perform those tasks. The greater their natural abilities, the higher the likelihood that they will be able to accede to higher, better and more challenging levels of education, *ceteris paribus*. Amid the informational asymmetries present in the labour market, particularly regarding levels of unobserved productivity, employers and clients can thus use individuals' positions

[7] Investments in human capital yield not only private returns, but also a variety of other static, dynamic, and non-pecuniary spill-over effects that may benefit society as a whole. Sianesi and van Reenen (2003, p. 161) argue, for instance, that the expansion of individual human capital stocks not only augments the productivity of neighbouring factors of production and technological processes, but may also lead to better public health, citizenship, and social cohesion. While the existence of social rates of return to human capital investments are acknowledged, the focus in this study falls exclusively on the private returns as they accrue to the individual and manifest in the labour market.

[8] In the remainder of this paper, *educational attainment* specifically refers to the number of years of education which an individual has successfully completed.

[9] The Sorting Hypothesis encompasses the theories of signalling, screening, sheepskin effects and credentialism. For a comprehensive discussion of this hypothesis and the relationships and interplay between its underlying theories, see (Weiss, 1995).

in the educational attainment distribution to probabilistically draw inferences about their expected productivity levels (Spence, 1973, p. 360). By implication, even if education served no human capital augmenting function, it would still appear to yield labour market returns because of its signalling function.

The SH is often mistakenly seen as an attempt to discredit the assertions of the HCT. However, the SH merely contends that human capital is partially innate and that part of education's function in the labour market is therefore purely informational.[10] In fact, while human capital, as it is defined in Section 2.1, is rooted in the HCT, it is also coherent with the SH. Both theories are consistent with observing positive returns[11] to education and under both theories, the returns that are observed are returns to underlying human capital. The primary difference is that under the HCT education augments productivity and under the SH education reflects innate productivity. This makes the HCT vs SH debate largely irrelevant at the level of the individual, since investment in education remains profitable irrespective of which theory is more pertinent.[12] However, the two theories do provide theoretical justification for the fact that education is, at the very least, a valid proxy indicator of human capital (Kingston *et al.*, 2003, p. 55).

The positive association between educational attainment, the probability of being employed, and labour market earnings is well-established in the international literature. In a comprehensive survey of more than 40 years of micro research on education-earnings linkages, Psacharopoulos and Patrinos (2002) conclude that, while structural shifts in economies and technological advances have altered the types of labour that are generally demanded, there remains compelling evidence that investments in educational attainment unambiguously yield private labour market returns in terms of improving both the employment and earnings prospects of labour force participants (Blundell *et al.*, 1999, p. 18). There is also increasing evidence that the earnings returns to education are not only comparatively large in relation to the returns on other investments, but that they exhibit convexity in a large number of countries, increasing in magnitude as individuals progress upwards in the educational attainment distribution (Colclough *et al.*, 2008; Harmon *et al.*, 2003, p. 115).

The concurrent shortage of skilled workers and apparent excess supply of unskilled labour in the South African labour market is one of the most perverse outcomes of the racially inequitable

[10] It is true that the strong versions of the screening and signalling hypotheses contend that human capital is entirely innate and thus immutable by education. However, studies examining the empirical validity of the SH have found such stringent assertions to be largely unsubstantiated (Brown and Sessions, 1998, p. 587). In reality, education most likely performs both sorting and augmenting functions, with the relative contribution of each role to labour market returns being case and context specific (see Arabseibani and Rees (1998); Brown and Sessions (1998, 2006); Clark (2000); Castagnetti *et al.* (2005) for evidence of the SH in international labour markets and Koch and Ntege (2006, 2008) for evidence of the SH in South Africa.)

[11] Unless explicitly stated otherwise, the term "returns" is used throughout this paper to refer to labour market returns either in the form of an increase in earnings or a rise in the probability of procuring employment.

[12] This paper abstracts entirely from individual human capital investment decision processes and instead takes human capital stocks as given.

distribution of education under apartheid (Burger and Von Fintel, 2009; Mariotti and Meinecke, 2009, p. 1). Moreover, the existence of unemployment even among those at the upper end of the educational attainment distribution suggests that a large part of South Africa's education sector is failing to instil the type and quality of skills that are valued in the labour market (Pauw *et al.*, 2008, pp. 46-47). Given the extent of the apparent mismatch between labour supply and demand and the strong racial dimension of this mismatch, differential returns to education between race groups and convexity in the general structure of educational returns in South African are common empirical findings in the earnings function literature (Keswell and Poswell, 2004; Daniels, 2007, p. 29).[13]

Numerous studies have investigated the earnings returns to education in the South African labour market, producing marginal return estimates ranging from as low as 0% for primary schooling to as high as 100% for tertiary education (Mariotti and Meinecke, 2009, pp.1-2). Similarly, educational attainment is found to have a strong non-linear impact on the probability of being employed in South Africa (Keswell, 2004). Branson *et al.* (2009, p. 47) estimate that individuals who have completed secondary school and individuals who have completed some form of tertiary education are respectively between 30% and 60% and 200% and 300% more likely to procure employment in the South African labour market than individuals with less than secondary educational attainment levels. However, the marginal employment returns to educational attainment estimated by Leung *et al.* (2009) and Oosthuizen (2006), while still indicative of considerable convexity, are significantly lower than those posited by Branson *et al.* (2009).

The substantial variation in South African labour market returns to educational attainment estimates across different studies casts doubt on the reliability of any single set of existing point estimates of the returns to education in South Africa. In general, obtaining unbiased and robust estimates of the returns to educational attainment is already complicated by issues such as omitted variable bias and sample selection bias (Heckman, 1979; Parsons and Bynner, 2005). However, in the South African context an additional concern is vast differences in quality of schooling obtained in "formerly White" as opposed to "formerly Black" parts of the formal education system. These quality differentials imply that the labour market benefits of South African education remain highly unequally distributed across race (Casale and Posel, 2010; Leibbrandt *et al.*, 2005). In essence, the failure to account for this feature of the South African schooling system may result in a further education quality bias in the estimates of the labour market returns to educational attainment. It follows that any prudent analysis of the South African labour market returns to educational attianment should take explicit cognisance of the racial differentials in the quality of education, how these differentials translate into inequitable labour market outcomes between race groups, and how the variation in school quality may impact on the structure of the returns to education across the attainment distribution.

[13] See Keswell and Poswell (2004) for a summary of studies providing evidence of increasing marginal returns to educational attainment in South Africa.

2.2.2 Returns to School Quality

The mounting international evidence that labour market return structures to educational attainment may be convex in a large number of countries coupled with the realisation that attainment levels often fail to accurately reflect productive skill levels has over the past two decades resulted in the proliferation of the number of studies investigating the impacts of school quality on both educational attainment levels and labour market outcomes (Kingston *et al.*, 2003, p. 55).[14] In theory, school quality should function as a catalyst for the observed labour market returns to educational attainment irrespective of whether education performs primarily a human capital augmenting or human capital reflecting function. Insofar as the HCT holds and education serves to augment existing aptitudes and endow individuals with new labour market-relevant knowledge and skills, better quality education should, all else being constant, result in more learning, more skill formation, and more growth in productive capacities. By implication, one would expect individuals who have received better quality schooling to also receive higher rates of return to educational attainment in the labour market than individuals who received poorer quality schooling, *ceteris paribus*. However, this result should also obtain even if, in accordance with the SH, education merely performs a human capital signalling function. Just as an individual's level of educational attainment may act as signal of ability or productivity, so too the quality of that education, when it is observable, may serve to either undermine or reinforce the fidelity of the signal. If the competency-based tasks which must be completed in order to accede to higher levels of education in high-quality educational institutions are known or perceived to be more rigorous or of a higher standard than those in low-quality schools, educational attainment in those schools would reflect different levels of underlying human capital. Consequently, low-quality education would again be expected to result in lower labour market returns to educational attainment, *ceteris paribus*. Whether educational attainment thus serves as a signal or a human capital augmentation device, the quality of education should invariably influence its function as either.

Measurement of school quality is often conceptually problematic. The existing literature on school quality and the labour market has therefore implemented a number of different measures in an attempt to adequately capture school quality-labour market linkages. These measures include inputs such as pupil-teacher ratios, expenditure per pupil and school textbook endowments as well as output and performance measures such as average cognitive test performance scores. While earlier studies focussed primarily on input measures, there is increasing evidence that output indicators may more accurately reflect the quality of schooling in terms of the extent to which it influences labour market outcomes (Moses, 2011; UNESCO, 2004, p. 40). Nevertheless, school performance outcomes are unlikely to accrue from school quality alone.

[14] Unless stated otherwise, the term "school" is used throughout this paper to refer to a formal primary, secondary, or tertiary educational training institution.

Instead, they are intimately related to pupils' and students' familial backgrounds and abilities as well as other socio-economic factors that influence their surrounding schooling environment (Yamauchi, 2011, p. 147). Consequently, school outcome measures may, at best, only be crude proxies for school quality.

In addition to measurement issues, assessment of the impact of school quality on labour market outcomes is complicated by the causal relationship that exists between school quality and educational attainment. Card and Krueger (1996, p. 43), Hanushek *et al.* (2008, p. 69) and others have provided empirical evidence that educational attainment is, on average, positively influenced by school quality, with individuals in high-quality schools being less likely to drop out and more likely to accede to higher levels of attainment than individuals in low-quality schools. This causal relationship has also been observed in the South African education system. Case and Yogo (1999, p. 3) find that improvements in school quality as measured by reductions in pupil-teacher ratios in South Africa have had a significant and positive impact on individuals' educational attainment levels.

Given the reinforcing effect that school quality has on the labour market impacts of educational attainment, the causal relationship between these two factors may in theory serve to explain the convex shape of the returns to education which is now commonly observed in developing countries. If individuals at the upper end of the attainment distribution are also likely to have attended high-quality schools they would be expected to earn a quality-premium on their labour market returns to education. However, in the absence of an explicit measure of school quality this premium would appear to accrue exclusively from high levels of educational attainment.

The causal relationship between school quality and educational attainment has a uniquely racial dimension in South Africa. As mentioned above, South Africa's education system remains characterised by large differentials in the quality of schooling received by different race groups. Rooted in the discriminatory education policies of the Apartheid-era, these differentials have persisted largely due to the governments failure to significantly improve the quality of education in "formerly Black" schools (Yamauchi, 2011, p. 148). As such, the average racial educational attainment differentials observed in South Africa are a partial reflection of racial differentials in school quality. While this has strong implications for the interpretability and generality of results form labour market returns to education estimations in South Africa, it may moreover provide an explanation for the persistence of racial labour market outcome differentials in the country. If, on average, Whites attend better quality schools than Blacks, it should be expected that they would face superior labour market outcomes even after controlling for educational attainment levels. This would mean that the unexplained component of the racial gaps in employment and earnings outcomes in South Africa, often perceived to be the result of persistent labour market discrimination, may actually be the result of pre-labour market discrimination in terms of the inequitable provision of access to good quality schooling.

The aforementioned hypothesis finds support in a number of studies. In one of the earliest published papers on the effects of school quality in the South African labour market, Moll (1992, p. 8) estimates a significant improvement in the earnings returns to educational attainment for Blacks during the sixties and seventies as a result of an improvement in education quality. Similarly Pillay (1994) finds that the poor quality of Black education in South Africa may undermine employers' confidence in the human capital signal sent by Black educational attainment credentials, thus resulting in differences in the rates of return to Black and White educational attainment. In a more recent study, Burger and Van der Berg (2011), using historical matric data in conjunction with Labour Force Survey data, estimate the "school quality component" in standard labour market discrimination measures and find that the variation in their proxy for school quality accounts for a substantial portion of the Black-White earnings gap in South Africa.

The discussion above suggests at least three important reasons for investigating the returns to school quality in South Africa. First, given the international evidence on the importance of school quality for labour market outcomes it is important to evaluate the relative contributions of the quantity of educational attainment and the quality of that attainment to employment and earnings prospects in South Africa. An understanding of these relative contributions is necessary in order to identify whether policy interventions aimed at increasing attainment levels or increasing the quality of education would be most effective in improving the socio-economic outcomes faced by South Africans. Second, the inclusion of a measure of school quality in returns to education estimations could provide insights as to the reason for the strongly convex structure of the returns to education in South Africa. Finally, the assessment of the labour market returns to school quality may allow one to gauge the extent to which racial differences in labour market outcomes in South Africa are driven by pre-labour market inequalities as opposed to labour market discrimination. These three hypotheses are investigated in the empirical analyses in Chapters 5 and 6.

2.2.3 Returns to Numeracy

It is commonly acknowledged that numeric ability is intimately related, though not necessarily commensurate, to overall cognitive ability.[15] As a result, the use of numeric competency test scores as a proxies for cognitive ability in labour market outcome estimations has grown rapidly in the international literature on the labour market returns to human capital. However, while numeracy may be correlated with general intelligence, there are indications that it also

[15] For the purposes of this study, numeracy may be defined as the extent of one's capacity to utilize and apply mathematical techniques, logic, and reasoning along with underlying mathematical principles in a functional manner, both in terms of solving mathematical problems and in terms of analytically assessing and solving non-numeric problems.

has an independent impact on various labour market outcome-related behaviours. Using several framing studies, Peters *et al.* (2006, p. 413) find that, when faced with complex tasks, individuals with greater numeric abilities extract relevant information faster, make superior judgements and decisions, and are less susceptible to irrelevant external influences than individuals with lower levels of numeracy, even after controlling for intelligence levels.[16] Similarly, Couper and Singer (2009, p. 17) find that numeracy, in conjunction with literacy, plays a critical role in the understanding and assessment of risks and the ways in which individuals deal with those risks. These associated capacities imply that numeracy may raise labour market productivity, not only because of the strong positive association it shares with general intelligence, but also because numeric skills are integral to the performance of labour (Wedege, 2002, p. 23).

McIntosh and Vignoles (2001, pp. 453-454) emphasise numeracy, alongside literacy, as one of the most basic and essential skills necessary to function in modern-day labour markets. Controlling for educational attainment and family background, the authors find statistically significant and large earnings returns even to basic numeric competency in Britain. Using data from two British panel surveys, Parsons and Bynner (2005, pp. 4-7) further show that numeracy is at least as important as literacy for success in the labour market and that individuals with low levels of numeracy are not only less likely to progress to higher levels of educational attainment, but also have poorer employment and earnings prospects than those with high levels of numeracy. Similarly, Rivera-Batiz (1992, pp. 325-326) finds that numeracy has a significantly large and positive impact on the probability of procuring employment, even after controlling for educational attainment levels and other indicators of human capital. Numeracy skills may thus serve as a hedge against unemployment, particularly for historically under-represented groups in the labour market, including females, Blacks, and the youth (Steen, 1990, p. 227).

The above-mentioned findings affirm the notion that numeracy is not only an important component of cognitive ability, but that it is also a component which may be independently valued in the labour market. This value is recognized in policy circles: the South African government has identified numeracy as one of the most critical and demanded skills in the South African labour market (Department of Labour, 2003; Daniels, 2007, p. 2). Despite the importance of numeracy, however, surprisingly little empirical research exists on the extent to which numeracy skills may influence labour market outcomes in South Africa.

Using data from the Project for Statistics on Living Standards and Development (PSLSD), Moll (1998, p. 289) finds significant earnings returns to numeric ability in South Africa and argues that, in addition to other inequitable outcomes produced by an historically segmented education system, pervasive differentials in numeracy levels along sociodemographic dimensions may be a strong underlying determinant of the inequitable distribution of labour market earnings in the country. In a more recent paper, Lam *et al.* (2008, p. 29) use data from the Cape Area Panel

[16] Peters *et al.* (2006) use Standardized Aptitude Test (SAT) scores to proxy for intellegence.

Study (CAPS) to examine the impact of numeracy, literacy and educational attainment on youth employment outcomes in South Africa. The authors find that numeracy and quantitative literacy levels are also strongly and positively related to the probability of being employed. These findings offer some preliminary support for the government's claim that numeracy constitutes a priority skill in the South Africa labour market. However, in order to gain an understanding of the importance of numeracy in relation to other components of human capital, more research is necessary.

In addition to the finding that numeracy may be an important determinant of both earnings and employment prospects, estimates of the employment and earnings returns to educational attainment are likely to be influenced by explicitly controlling for numeracy and literacy measures. Charette and Meng (1998, p. 516), for example, find that the earnings returns to educational attainment may be upward biased by as much as 20% if numeracy measures are excluded from returns estimations. Similarly, Lam *et al.* (2008, p. 29) show that estimates of the employment returns to schooling in the Cape Town metropolitan area may drop by up to 50% and even become statistically insignificant when one explicitly controls for numeracy and literacy levels.[17] Understanding the role of numeracy in the South African labour market is therefore important for at least two reasons. First, to the extent that numeracy is reflective of ability, its inclusion in labour market returns regressions may serve to mitigate the magnitude of the bias in education return estimates which would otherwise arise from omission of a direct measure of ability. Consequently, controlling for numeracy should allow for a more nuanced analysis of the labour market returns to educational attainment in South Africa. Second, given the explicit emphasis on the value of numeric skills in the South African labour market, directly estimating the returns to numeracy may provide more definitive indications of its labour market value relative to other human capital measures, including educational attainment and school quality.

[17] Since the CAPS data used in Lam *et al.* (2008) covers only youths and young adults in metropolitan Cape Town, this result is unlikely to hold exactly for the greater South African population of working-age. However, it does provide some indication of the importance of both numeracy and literacy for the probability of procuring employment in South African metropolitan areas.

Chapter 3

Data Features and Sampling Considerations: The National Income Dynamics Study (NIDS)

Because of the limitations on historically available micro-level data, few studies have attempted to provide an integrated and cohesive empirical analysis of the manifold private labour market returns to various human capital proxies. Such an analysis requires data that not only contains accurate information on multiple human capital and labour market variables, but also allows for observed individual labour market outcomes to be linked to human capital holdings.

The 2008 National Income Dynamics Study (NIDS) is one of the first datasets to contain concurrent information on individual labour market outcomes, educational attainment levels, numeric proficiency and quality of schooling received in South Africa.[1] Given the presence of these and various other sociodemographic and human capital-related variables, the data seems potentially suited for the type of analysis outlined above. However, while the 2008 NIDS dataset is rich in its coverage, it is also characterised by limited and selective response patterns on many key variables that need to be used in order to obtain valid estimates of the labour market returns to human capital. While these response patterns are in themselves interesting grounds for scientific inquiry[2], they constitute potential sources of estimation bias that need to be accounted for when analysing private human capital returns. Before proceeding with the main analysis, it is therefore appropriate to first give an overview of some key features of the NIDS data. The

[1] Unless stated otherwise, *NIDS* and *NIDS 2008* are used interchangeably throughout this paper to refer to Wave 1 (2008) of the National Income Dynamics Study.

[2] Du Rand *et al.* (2010), for example, provide an extensive analysis of the underlying nature of the response patterns to the NIDS numeracy test module.

following sections outline aspects related to response rates and sampling in the data, with specific emphasis on the nature of, and response patterns to, the NIDS *numeracy* and *school quality* variables.

3.1 NIDS Background and Sampling Considerations

The National Income Dynamics Study (NIDS) is the first nationally representative longitudinal household survey in South Africa.[3] The primary purpose of the study is to investigate the dynamics of household structures and the changes in household welfare and well-being in South Africa by examining incomes and expenditures, labour market outcomes, asset holdings, health, education and other dimensions of socioeconomic welfare. (Leibbrandt *et al.*, 2009*b*, p. 1)

Wave 1 of NIDS was enumerated by the Southern African Labour and Development Research Unit (SALDRU) in 2008 and surveyed a total of 28 255 individuals from 7 305 households.[4] Of the 18 639 adults (15 years or older) included in this sample, 1754 were unavailable at the time of the survey interview and had to have proxy questionnaires completed on their behalf by other household members. A further 1 246 adults refused to complete the adult questionnaire section. These individuals therefore have missing data on many of the labour market and human capital items that were only documented in the NIDS adult questionnaire. (NIDS, 2009; Leibbrandt *et al.*, 2009*a*,*b*, p. 22)

Overall, 83% of the eligible sample responded to the NIDS adult questionnaire. However, many of the labour market outcome, demographic, and human capital indicators that were captured via this questionnaire were subject to considerable item non-response.[5] Table 3.1 below provides a breakdown of the sample sizes and number of non-missing observations on some of the key variables that are used in the empirical analyses in Chapters 5 and 6. Since the estimation of labour market returns is the primary focus of this paper, the analysis that follows only considers those individuals in the population of working age (15 to 65 year-olds).[6] The breakdown of sample sizes within this group is therefore presented in columns 4 and 5 alongside the breakdown for the full NIDS sample in columns 2 and 3.

[3] A NIDS panel will ultimately be constructed from the various waves of NIDS that are enumerated every two years. However, only the first wave of the data is currently available.

[4] While a total of 31 170 individual household members were identified in the 7 305 participating households, 2 915 non-resident household members (i.e. members who usually reside at the household for fewer than 4 nights a week) were excluded from the study in order to avoid double counting. This exclusion effectively limited the survey sample to 28 255 observations. (NIDS, 2009, p. 8)

[5] The literature on survey design distinguishes between two main types of survey nonresponse. Unit nonresponse occurs when a unit (normally an individual or a household) in the eligible survey sample fails to respond to any of the items in the survey questionnaire. By contrast, item nonresponse occurs when a unit fails to respond only to certain survey items, whether they be specific questions in the questionnaire or subsections of questions (Gilley and Leone, 1991, p282).

[6] Unless stated otherwise, all of the analyses that follow in Chapters 5 and 6 are conducted for the population of working-age only.

Given South Africa's historical context , it is often of interest to disaggregate labour market analysis by race group. However, the racial sampling used in NIDS may limit the scope for doing so. In the 2008 dataset, Coloured individuals are over-represented and White individuals under-represented relative to their actual population shares. This sampling discrepancy is visible in both the full and working-age survey samples and survey sampling weights have therefore had to be adjusted in order to ensure that reliable inferences about the South African population could still be drawn from the data. While the table thus shows that Coloured and White respondents respectively represent 15.61% and 5.86% of the working-age survey sample, their corresponding weighted population shares (using NIDS sampling weights) amount to 9.35% and 10.08%. As a result of under-sampling, the working-age sample includes only 974 White respondents and 294 Asian[7] respondents.[8] These small racial sub-sample sizes coupled with the fact that some of the respondents concerned may also have missing data on variables that are used in the labour market returns estimations could imply that there is not sufficient variation in the data to allow for identifiable parameters in separate within-group estimations.

As indicated in Table 3.1, the 16 627 working age respondents in NIDS constitute 58.85% of the total survey sample. While 30% of the 9 273 labour force participants[9] in this sample were either actively searching for work or discouraged work seekers, 70% indicated that they were formally, casually, or self-employed. However, only 5 765 (88.82%) of the employed respondents provided non-zero monthly earnings data and thus satisfy the fundamental prerequisite for inclusion in semi-logarithmic earnings function estimations.[10] These individuals therefore constitute the base sample for the empirical analysis of the earnings returns to human capital in the chapters that follow and are hereafter simply referred to either as *"earners"* or individuals in the *"earnings sample"*.

Within the group of earners there was a significant extent of non-response on many important work-related correlates of labour market earnings. For example, the table shows that only 92.63% and 82.64% of earners respectively provided information on the nature of the main type of occupation from which they derive their earnings and the number of hours they usually work on this job in an average week. Similarly, only 88% provided information on whether they belonged to a labour union or not - a factor which has been shown to have a significant impact on labour market earnings in South Africa (Azam and Rospabé, 2007, p. 421). Including these variables as covariates in an earnings function regression will thus reduce the size of the

[7] The Asian racial classification used throughout this paper also includes individuals of Indian decent.

[8] While the small sample size of the Asian race group is also partially the result of under-sampling, it is primarily the consequence of the relatively small scale of the NIDS survey (when compared to previous nationally representative household surveys) and the fact that Asians only constitute a small proportion of the South African population.

[9] The broad definition of the labour force is used throughout this paper.

[10] It is common practice to specify the dependent variable in Mincerian-type earnings functions in semi-logarithmic form (i.e. using the *log of earnings* as the dependent variable) since this allows the parameter estimates to be interpreted as percentage changes in earnings corresponding to unit changes in the covariates.

Table 3.1: Sample Sizes and Non-missing Observations in NIDS 2008

	Full Sample	**% of Sample**	**Age 15-65**	**% of Sample**	**Age 15-65 / Full Sample**
Total Sample	28255	100	16627	100	58.85
Black	22157	78.42	12721	76.51	57.41
Coloured	4166	14.74	2595	15.61	62.29
Asian	439	1.55	294	1.77	66.97
White	1432	5.07	974	5.86	68.02
NEA	17095	60.5	5792	34.83	33.88
Labour Force	9598	33.97	9273	55.77	96.61
Labour Force *(Base Sample: Labour Force Participants)*					
Unemployed	2814	29.32	2782	30	98.86
Discouraged	976	10.17	954	10.29	97.75
Searching	1838	19.15	1828	19.71	99.46
Employed	6784	70.68	6491	70	95.68
Employment *(Base Sample: Employed)*					
Casual	729	10.75	707	10.89	96.98
Self-Employed	874	12.88	834	12.85	95.42
Non-zero Earnings	5913	87.16	5765	88.82	97.5
Earners *(Base Sample: Non-zero Earnings)*					
Occupation	5459	92.32	5340	92.63	97.82
Hours Worked	4820	81.52	4741	82.24	98.36
Union Data	5165	87.35	5073	88	98.22
Human Capital Variables *(Base Sample: Total Sample)*					
Education	25146	89	16532	99.43	65.74
Numeracy Score	4353	15.41	3504	21.07	80.5
School Quality Score	4861	17.2	4759	28.62	97.9
Human Capital Variables *(Base Sample: Non-zero Earnings)*					
Education	5811	98.27	5735	99.48	98.69
Numeracy Score	1001	16.93	1001	17.36	100
School Quality Score	1715	29	1708	29.63	99.59

NOTES: Figures are unweighted and thus correspond to the number of non-missing observations in the NIDS dataset. Figures may not sum to totals because of missing observations. Columns 3 and 5 show the number of non-missing observations on column 1 variables/samples as a percentage of the number of non-missing observations on the indicated base sample/variable for that section of the table. E.g., in the population of working age $\%_{Unemployed} = N_{Unemployed}/N_{Labour\ Force\ Participants} = 2782/9273 = 30\%$. Column 6 shows the number of non-missing working age observations on each of the variables/samples in column 1 as a percentage of the number of non-missing observations on those variables/samples in the full survey sample.

estimation sample.[11]

The potential for estimation sample censoring due to the inclusion of covariates with a large proportion of missing observations is not in itself necessarily a cause for concern. Provided that the extent of censoring is not too severe and that the pattern of missing values on covariates is either randomly determined or can be defined completely in terms of the variation in other observables, the estimation sample may still have sufficient variation and be representative enough to allow for unbiased estimation of labour market returns. However, as is indicated in Table 3.1, the small number of non-missing observations on the two human capital measures *numeracy* and *school quality* imply that sample sizes could be dramatically reduced in estimations that include these variables as regressors. While the highest level of educational attainment was successfully documented for nearly all earners, only 1708 (29.63%) have any data on the quality of the secondary schooling they received and only 1001 (17.36%) have scores that reflect their level of numeracy. The extent of non-response on these two survey items makes it unlikely that the pattern of missing observations on the *numeracy* and *school quality* will be missing at random (MAR).

Educational attainment, school quality and numeracy constitute the three key human capital indicators that are the focus of the empirical analysis in this paper. However, it should be clear from the discussion above that it would be imprudent to simply disregard the negative implications that censoring on these variables may have for the reliability of human capital return estimates. Before proceeding with the descriptive and multivariate analyses, it is therefore important to first consider the respective patterns of non-missing observations on the numeracy and school quality variables and evaluate what implications these patterns may have for the ability to produce unbiased estimates of the labour market returns to human capital in South Africa.

3.2 The NIDS Numeracy Score

NIDS 2008 is one of the first datasets to incorporate a numeracy test module in a general household survey. An advantage of this feature is that the data may allow numeracy levels to be connected to individual labour market outcomes. However, since the primary aim of NIDS is to measure aspects directly related to household welfare rather than measuring individual performance on aptitude tests, the numeracy test never constituted a priority module for survey enumerators. In addition, the test was enumerated within households instead of an educational

[11] The extent of the reduction in sample size depends on the extent to which missing values on one covariate overlap with non-missing values of another. These overlaps are not shown in Table 3.1. It is therefore likely that the inclusion of the *union* dummy variable in a Mincerian earnings function, for example, will limit the estimation sample size to less than 5 073 observations.

environment and was targeted at individuals from a wide age spectrum. Many of these individuals may not have been used to or felt comfortable with taking cognitive assessment tests and consequently lacked the confidence and/or will required to participate in the module. When these issues are considered in light of the fact that participation in the test was voluntary and that respondents were given no explicit incentive to participate, it is perhaps not surprising that only 21% of working-age respondents decided to take the test.

3.2.1 The Nature of the NIDS Numeracy Test

The NIDS numeracy test was based on South Africa's national schooling curriculum and aimed to assess respondents' levels of numeric, algebraic, measurement, spacial, and data competency (Griffin *et al.*, 2010, p. 2). The test was enumerated in four different difficulty levels, each of which was targeted at individuals within a specific range of mathematical attainment. The level of the test that respondents were supposed to write therefore depended on the highest level of mathematics which they had completed at school.[12] Each test consisted of 15 intellectually challenging multiple choice questions. Respondents were given ten minutes to complete the test and it was emphasized that the results of the tests would remain confidential and would not be revealed to participants afterwards.[13] In order to account for the differences in difficulty between test levels and between the different questions in each test, the raw test scores were calibrated using item response modelling and a standardized numeracy score was constructed (Griffin *et al.*, 2010, p. 22). Subsequent consistency checks were also performed to ensure the nomological validity of the resultant numeracy score (Griffin *et al.*, 2010, p. 21). The numeracy variable in the NIDS data can therefore be interpreted as a valid indicator of individuals' abilities to successfully perform tasks that are numeric in nature.

3.2.2 The Representativeness of the NIDS Numeracy Score[14]

Hanushek and Woessmann (2010, p.4) suggest that a response rate of 85% should be used as a benchmark for reliability when considering the results from cognitive assessment tests. As shown in Table 3.1, however, the response rate to the NIDS numeracy test module for respondents in the population of working age is only 21.07%. With such a significant extent of non-response, it is highly unlikely that numeracy data will be missing-completely-at-random

[12] Despite this intended channelling, the data reveals that respondents were allowed to choose which level of the test they wanted to take. As a result, some respondents chose to write tests that were comparatively easy and others ones that were comparatively difficult relative to the tests that they were supposed to take. (Du Rand *et al.*, 2010, p. 4)

[13] Respondents who had not yet completed the test after 10 minutes had elapsed were granted a further 5 minutes writing time.

[14] For a comprehensive analysis of the NIDS numeracy test participation decision, see Du Rand *et al.* (2010).

Table 3.2: Summary Statistics for NIDS Samples with and without Numeracy Data

	Took Test		**Did Not Take Test**	
	N	*Mean*	*N*	*Mean*
Age	3504	26.15	13123	36.19
Years of Education	3495	9.96	13037	7.99
School Quality Score	1684	0.07	3075	0.08
	N	*%*	*N*	*%*
Male	1511	20.51	5857	79.49
Female	1993	21.53	7266	78.47
Black	2833	22.27	9888	77.73
Coloured	535	20.62	2060	79.38
Asian	18	6.12	276	93.88
White	117	12.01	857	87.99
Urban	1943	22.65	6636	77.35
Rural	1561	19.40	6487	80.60
Total	**3504**	**21.07**	**13123**	**78.93**

NOTES: Estimates are unweighted. The sample includes only respondents in the population of working age.

(MCAR) and that the numeracy score variable can be used in subsequent analysis without taking cognisance of the response pattern in the methodology.[15]

The numeracy test module not only constituted a particularly cognitively challenging item in the NIDS survey, but was also enumerated at the end of the survey questionnaire. Given that more cognitively challenging items cause greater response fatigue and that response fatigue is cumulative, it would have been rational for respondents to expect that participation in the test module would require a higher degree of effort than the majority of the other survey questions (Axinn and Pearce, 2006, p. 42). Since taking the test was voluntary and no explicit material incentive was offered in order to evoke participation, one would expect, *a priori,* that a respondent's propensity to participate in the test would largely depend on the degree of personal effort required to do so (Bradburn, 1978, p. 37).

Table 3.2 presents the sample sizes, means, and proportions of some of the key sociodemographic variables used in the empirical analysis in Chapters 5 and 6 for those respondents who took the numeracy test and those who did not. The differences in the various estimates for the

[15] In the context of non-response, data are MCAR when the probability of observing a response is completely independent of both observable and unobservable factors (Cobben, 2009, p. 10).

two sub-samples show that the NIDS numeracy data is certainly not MCAR. Test respondents are found to be younger on average and have higher levels of educational attainment than those who opted out of taking the test.[16] This is consistent with the argument given above since, *ceteris paribus*, younger respondents should find it easier to recall what they learned about mathematics at school or at university (Glazerman *et al.*, 2000, p. 20). Similarly, respondents who attained high levels of education are probably better prepared, and therefore potentially more inclined, to take a numeracy test than those who did not Chevalier *et al.* (2009, p. 718). The table also shows that there is considerable variation in response rates between the different race groups. While the response rates for Blacks and Coloureds are close to the overall response rate for the population of working age, only 12.01% of Whites and 6.12% Asians in the eligible sample took the test. These differences suggest that the NIDS numeracy score may not be a representative measure of the numeric ability of older respondents or those positioned lower down in the educational attainment distribution in South Africa.[17] Similarly, if the numeric ability distributions for Whites and Indians differ significantly from those for Coloureds and Blacks, the representativeness of the numeracy score will be biased in favour of the latter two race groups.

Given that it is possible to explicitly control for any observable factors that influence the numeracy test response decision and the numeracy test score outcome, the aforementioned sources of bias are largely ignorable. However, since the numeracy test response rate appears to be a function of at least one observable human capital measure, educational attainment, it is likely that it will similarly be correlated with certain unobserved human capital measures, including innate ability and intrinsic motivation.[18] In other words, it is possible that the response rate is subject to non-ignorable ability bias. To illustrate this, Table 3.3 presents the results from two fairly rudimentary earnings and employment returns regressions, estimated separately for the respondents who took the NIDS numeracy test and those who did not.[19] The results indicate that test respondents have, on average, both higher marginal earnings and higher marginal employment returns to educational attainment than their non-test-taking counterparts.[20] This finding suggests that test respondents may be those individuals who have higher abilities or who possess certain unobserved characteristics which enable them to capitalise on additional investments in education.[21] However, these unobserved factors are also likely to have a direct impact on

[16] The difference in average school quality for the two groups is also statistically significant at the 1% level.
[17] This is based on the assumption that average numeracy levels differ with age and educational attainment.
[18] This hypothesis would hold if respondents at least partly based their decisions to participate in the numeracy test on how well they anticipated to perform in the test, how difficult they expected the test to be, or how rewarding they expected the test-taking experience to be (Du Rand *et al.*, 2010, p. 8).
[19] These estimations do not control for any form of sample selection or other issues that may lead to bias in the estimates.
[20] The Wald tests for cross-equation parameter equivalence reveal that one may reject the hypothesis that coefficients on the educational attainment variable in the two earnings estimations and the two employment estimations are not different from one another at below 1% significance.
[21] It should be emphasized that, while differences in the returns to education are used here as an indication of ability or other similar underlying human capital differences, the observed results may also be driven by a multitude of

Table 3.3: Rudimentary Earnings and Employment Returns Estimations for Samples With and Without Numeracy Data

	Log of Earnings		**Pr(Employment)**	
	No Test	Took Test	No Test	Took Test
Age	0.093***	0.120***	0.209***	0.226***
Age^2	-0.001***	-0.001***	-0.002***	-0.003***
Education	0.120***	0.189***	0.049***	0.066***
Female	-0.488***	-0.526***	-0.539***	-0.591***
Coloured	0.096***	0.107	0.264***	0.346***
Asian	0.856***	0.673**	0.083	1.483***
White	0.916***	0.808***	0.422***	0.606***
Rural	-0.238***	-0.139**	-0.085***	0.060
Constant	4.407***	2.944***	-4.171***	-4.834***
Observations	4734	995	11922	3452
R^2	0.422	0.395		
F-stat	437.860	97.040		
P-value	0.000	0.000	0.000	0.000
Area under ROC curve			0.751	0.822

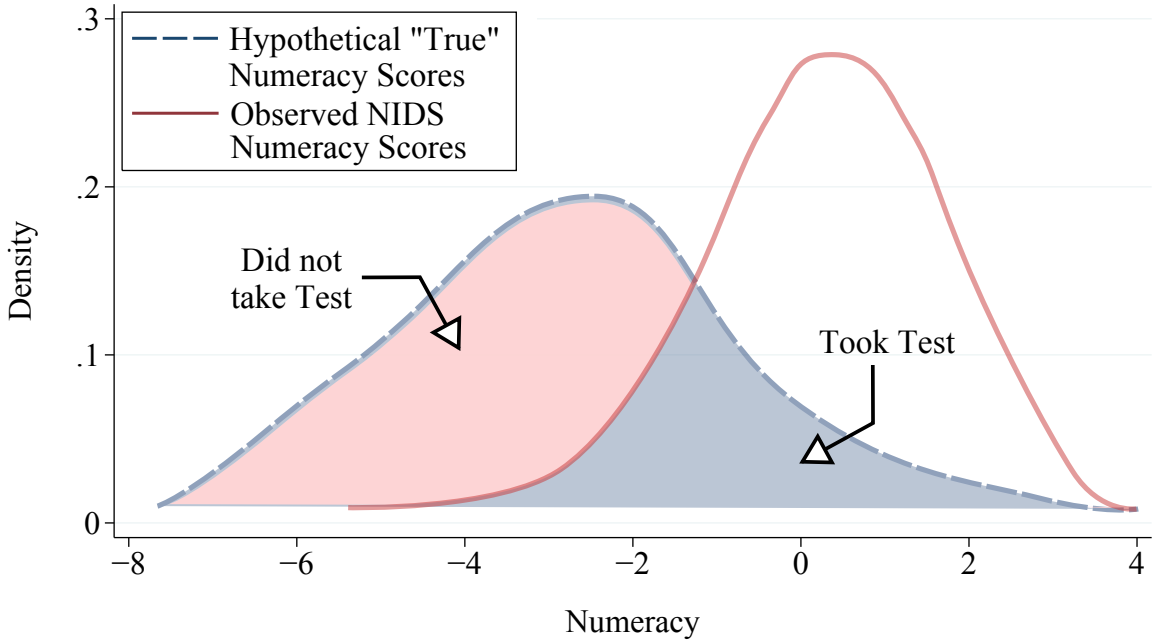
NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on robust standard errors. The dependent variable in the first two columns is the *log of monthly earnings* and the models are estimated via OLS. The dependent variable in the second two models is a binary variable indicating whether or not a respondent is employed and the models are estimated using probits. Reference category for the race dummies is Black. The *education* variable measures the number of years of completed formal education.

numeric ability and, therefore, on the numeracy scores that are observed. By implication, the NIDS numeracy measure will not be representative of the numeric abilities of those individuals who possess these unobserved characteristics in lesser amounts. Specifically, if it is the case that more able or more motivated individuals were also more likely to participate in the NIDS numeracy test, as is suggested here, the resultant numeracy measure would fail to adequately capture the bottom tail of the ability distribution. As a result, the numeracy score would not only exaggerate average numeracy levels, but also understate the extent of the variation in numeric abilities between individuals and groups. This notion is illustrated conceptually in Figure 3.1 which shows the observed distribution of the NIDS numeracy scores against a hypothetical "true" distribution of numeric ability in the South African population of working age.

The low response rate on the NIDS numeracy test coupled with the fact that the response pattern

other unmeasured factors.

Figure 3.1: Numeracy Distributions by Race Group



NOTES: The kernel density for the NIDS numeracy score was calculated using the Epanechnikov kernel with a bandwidth of 1. Results are unweighted. The curve for the "true distribution is completely hypothetical and artificial.

appears to be a deterministic function of both observed and unobserved human capital correlates complicates unbiased estimation of the South African labour market returns to numeracy for several reasons. First, as illustrated in Figure 3.1, the nature of the non-random selection into the test implies that the resultant numeracy score may be an upward-biased and left-skewed estimate of the true distribution of the numeric abilities of the population of working-age. Failure to fully account for this fact in the estimation analysis would therefore bias the labour market returns to numeracy estimates downwards. Second, any significant reduction in estimation sample size as a result of using the numeracy score as an explanatory variable could adversely impact on the precision with which the parameters of interest can be estimated. Third, and most importantly, if any of the unobserved factors that are associated with the numeracy test participation decision share similar associations with labour market outcomes, the use of the NIDS numeracy score as an explanatory variable in labour market returns estimations would result in sample selection bias. This is because the observation of the dependent variable for any regression where the numeracy measure is used as a covariate is conditional on the observation of a numeracy score or, put differently, on the probability of having participated in the numeracy test module. If the residual component of this test participation probability includes unobservables that are also present in the residual for the labour market outcome estimation, the parameter estimates for that outcome will be biased. It is therefore necessary to explicitly control for the non-random probability of observing the NIDS numeracy score when trying to estimate the labour market returns to numeracy. This issue is discussed in more depth in Chapter 4.

3.3 School Quality Data in NIDS

NIDS 2008 is particularly rich in its coverage of education information. Among a multitude of other schooling-related questions, respondents who answered the adult questionnaire were asked to provide details on the educational institution where they completed their highest grade of formal schooling or, if they were still in school, on the educational institution where they were currently enrolled. To preserve respondent confidentiality and anonymity, this information was not made available in the general release of the NIDS 2008 data. However, the data that was collected allowed for some of the respondents to be successfully linked to schools in the South African schools registry. Using the 2008 National Senior Certificate Examinations (NSCE) results, various composite indicators of school output performance were then created for each of the schools that could accurately be identified. This made it possible to allocate to each of the successfully matched respondents in the survey a set of school quality indicators. Among these indicators is a standardized measure of the aggregate average mark which the respondent's school achieved in the 2008 SCE.[22] It is this variable that is used to proxy for school quality in the analyses that follow. As such, the variable is referred to as the *NIDS school quality score* throughout this study.

3.3.1 The Nature of the NIDS School Quality Measure

The availability of a school quality measure in the NIDS data potentially allows for a much more nuanced analysis of the South African labour market returns to human capital. However, before the measure is used in the analysis, it is important to consider its various advantages and shortcomings. On the plus side, the literature on the labour market returns to school quality suggests that output measures such as school performance may provide better indications of school quality and stronger links between school quality and labour market outcomes than input measures like teacher-pupil ratios or mean expenditure per pupil (Hanushek, 2005; Moses, 2011). However, the NIDS quality variable directly captures average school performance only at the highest level of secondary schooling, i.e. Matric. This has at least three significant implications. First, it means that school quality is only observed for those individuals who attained at least some secondary schooling. Thus, the quality of schooling received is unknown for individuals with below-secondary levels of attainment. Second, the measure ignores the fact that individuals who received the same or similar quality schooling during secondary education, may have been exposed to very different quality education at the primary and/or tertiary level.[23] Third,

[22] To preserve the anonymity of those schools that could be identified and prevent "reverse engineering" of the school performance indicators, the standardized Matric average score was converted into a categorical variable with several bins.

[23] One could argue that the limited upward mobility of the majority of the South African population, coupled with the demanding entry requirements at better-performing educational institutions, make it likely that individuals

average NSCE results are likely to be upward-biased indicators of overall secondary school performance since the sample of individuals who manage to progress to Matric and sit the SCE are not a random group. The literature on school drop-out and repetition rates in South Africa suggests that below-average performing students are far more likely to drop out of school before reaching Matric than others (Motala, 1995; Panday and Arends, 2008). Drop-out rates are also higher in schools that have historically been poorer-resourced and performing at below average levels (Gustafsson, 2011; Burger and Van der Berg, 2011; Motala, 1995, p. 10). By implication, the upward bias in any school quality measured which is based on a school's overall SCE performance is, on average, expected to be inversely proportional to the quality of education in that school, *ceteris paribus*. It is therefore likely that the lower tail of the NIDS school quality distribution will understate the extent of poor quality schooling in South Africa.

The fact that the NIDS school quality variable measures school performance only in 2008 may be problematic given that only a small proportion of the NIDS respondents would have completed Matric in 2008. While it may be reasonable to assume that the quality of schooling provided in any given educational institution changes only slowly over time, it is not necessarily the case that individuals who attended the same school ten or twenty years apart received the same quality of education. As a result, the representativeness of the NIDS school quality score is likely to be biased in favour of younger respondents who left school more recently than older individuals. The quality score variable also does not control for the fact that some individuals may have switched schools during secondary education. There is no guarantee that the schools to which respondents could be matched are the same ones that they attended in the year prior to completing their indicated highest grades of attainment or where they received the bulk of their high school education. Since it is reasonable to expect that the effects of school quality will be compounded over the period of exposure to a school, this implies that the average quality of secondary schooling which some respondents would have received may differ from what is indicated by the average SCE results of the schools to which they were matched. [24]

Collectively, the above-mentioned issues imply that caution should be taken when interpreting the results from any analysis which employs the NIDS school quality score as an indicator of the quality of schooling in South Africa. While the variable is thus referred to and discussed as school quality throughout this study, it should be borne in mind that it is, at best, an imperfect proxy for the quality of secondary schooling in South Africa.

---

would receive similar quality schooling during primary, secondary, and tertiary education.

[24] However, it is acknowledged that this effect would be negligible on the overall representativeness of the NIDS school quality measure, unless a significant proportion of the respondents who could be matched to schools attended more than one school over the course of their secondary education and the quality of those schools differed systematically from one another.

3.3.2 The Representativeness of the NIDS School Quality Score

In addition to the issues mentioned above, Table 3.1 shows that the NIDS school quality score is observed for only 28.62% of working-age respondents.[25] There are several reasons for the large number of missing values on this variable. First, the questions pertaining to details of the schools where respondents completed their highest grades of formal education were, by definition, not applicable to the 1 615 individuals who indicated that they never received any formal schooling. Second, even if individuals with primary education as the highest indicated level of attainment provided information on the school which they attended, it would only have been possible to match them to school quality data if that primary school was paired with a secondary school. As a result, the NIDS school quality score is observed for only 229 of the 3 810 respondents with primary educational attainment levels.

The missing NIDS school quality scores for the two above-mentioned groups is largely a consequence of the nature of the school quality variable and the way in which it was constructed. However, the rest of the missing observations on this variable may be attributed directly to respondent behaviour. As was the case with the participation in the NIDS numeracy test module, some respondents may have refused to answer questions related to the schools which they attended. Others may have answered the questions, but failed to provide sufficiently accurate or comprehensive enough information to enable successful matching to a school in the South African schools registry. The missing values on the school quality measure for individuals with higher than primary educational attainment is thus a function of both non-response and inadequate response. While it is not possible to distinguish between these two sources of missingness in the data, both imply that the school quality data is unlikely to be MCAR.

Table 3.4 presents the sample sizes, means, and proportions of the same sociodemographic correlates of non-response that are shown in Table 3.2, for the sample of individuals for whom school quality is observed and the sample for whom it is not. The estimates in the table show that, similar to the case for participation in the numeracy test module, respondents who could be matched to school quality data are younger, have higher levels of educational attainment, and have higher levels of numeracy, on average, than those for whom school quality data is missing. Based on the information provided above, this is to be expected. However, even when the estimation samples are restricted to exclude any individuals with less that secondary educational attainment, the differences between the average age and years of completed education for the two sub-samples remain statistically significant at 1%. The differences in school quality match rates[26] between the different race groups, while less pronounced than the differences in the numeracy test response rates, are nevertheless statistically significant at the 1% level.

[25] If respondents who indicated that they had only primary levels of educational attainment or received no formal schooling are excluded from the sample, this estimate rises to 40.44% of working-age respondents.

[26] Here, "match rate" refers to the percentage of a sample that could successfully be matched to a school.

Table 3.4: Summary Statistics for NIDS Samples with and without School Quality Data

	School Quality		**No School Quality**	
	N	*Mean*	*N*	*Mean*
Age	4759	28.73	11868	36.22
Years of Education	4747	10.64	11785	7.51
Numeracy Score	1684	-0.45	1820	-0.55
	N	*%*	*N*	*%*
Male	1943	26.37	5425	73.63
Female	2816	30.41	6443	69.59
Black	3605	28.34	9116	71.66
Coloured	729	28.09	1866	71.91
Asian	95	32.31	199	67.69
White	328	33.68	646	66.32
Urban	2721	31.72	5858	68.28
Rural	2038	25.32	6010	74.68
Total	**4759**	**28.62**	**11868**	**71.38**

NOTES: Figures here are unweighted; in the regression analysis sampling weights are used to correct for this.

The estimates in Table 3.4 illustrate that the group of respondents for whom school quality is observed (hereafter called the *school quality sample*) is characteristically distinct from the rest of the NIDS working-age sample in terms of roughly the same observables as those that distinguish the NIDS numeracy test participants from the non-participants. Given the nature of these correlates, it is again likely that the respondents in the school quality sample will also differ from the rest of the working-age sample in terms of certain unobservable characteristics which may in turn be correlated to school quality. For example, since drop-out rates in South Africa have been found to be inversely related to school quality, one could infer from the higher average educational attainment levels of the school quality sample that these individuals may generally have attended better quality secondary schools than the rest of the working-age sample. If this were the case, it would imply that the NIDS school quality measure is subject to selection bias. Specifically, it would mean that the same unobserved factors which are associated with better school quality are also associated with a higher probability of observing the NIDS school quality measure.

To test the plausibility of the hypothesis that respondents in the school quality sample may be different from the rest of the sample in terms of unobservable characteristics which are related to school quality and/or labour market outcomes, the rudimentary earnings and employment

Table 3.5: Rudimentary Earnings and Employment Returns Estimations for Samples With and Without School Quality Data

	Log of Earnings		**Pr(Employment)**	
	No Quality	Quality	No Quality	Quality
Age	0.091***	0.111***	0.199***	0.256***
Age^2	-0.001***	-0.001***	-0.002***	-0.003***
Education	0.118***	0.223***	0.046***	0.111***
Female	-0.523***	-0.399***	-0.563***	-0.510***
Coloured	0.088**	0.120**	0.258***	0.371***
Asian	0.909***	0.787***	0.059	0.440***
White	0.959***	0.767***	0.486***	0.384***
Rural	-0.220***	-0.194***	-0.031	-0.126***
Constant	4.464***	2.767***	-3.988***	-5.708***
Observations	4030	1699	10677	4697
R^2	0.421	0.394		
F-stat	369.038	150.318		
P-value	0.000	0.000	0.000	0.000
Area under ROC curve			0.748	0.822

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on robust standard errors. The dependent variable in the first two columns is the *log of monthly earnings* and the models are estimated via OLS. The dependent variable in the second two models is a binary variable indicating whether or not a respondent is employed and the models are estimated using probits. Reference category for the race dummies is Black. The *education* variable measures the number of years of completed formal education.

outcome regressions presented in Section 3.2.2 are re-estimated for the school quality sample and the sample without the school quality variable. The results of these estimations are shown in Table 3.5. The Wald tests for cross-equation parameter equivalence of the coefficient estimates on the educational attainment variable shows that respondents in the school quality sample have significantly higher marginal earnings and employment returns to education, on average, than those for whom school quality data is missing. From these fairly simple estimations, it would therefore appear as though the school quality sample respondents possess certain attributes or capacities that make them more effective at converting increases in educational attainment into improvements in labour market outcomes.

The fact that selection into the school quality sample may be endogenous has similar implications for the analyses that follow as those discussed for the numeracy score sample in the previous section. First, the nature of the NIDS school quality measure implies that it will be

an upward-biased indicator of the quality of schooling which individuals have received. In addition, if individuals who attended better quality schools were also more likely to be matched to school quality data, as is suggested to be the case, this would result in a further upward bias in the NIDS school quality measure. As a result, one would expect the labour market returns to school quality estimates to be biased downwards, *ceteris paribus*. However, an even bigger problem arises from the fact that the unobserved correlates of the probability of observing the school quality measure are potentially correlated with labour market outcomes. If this is really the case and the school quality measure is used as an explanatory variable in labour market outcomes estimations, the observation of the dependent variable (earnings or employment) would be censored in terms of both the observable and unobservable factors that influence labour market outcomes. In other words, the use of the school quality measure in estimating either the labour market returns to school quality or the labour market returns to education when controlling for school quality in South Africa, would cause additional selection bias in the estimates. The results from the very preliminary analysis in this section thus suggests that it may be necessary to explicitly control for selection into the school quality sample whenever it is used as a covariate in the earnings and employment outcome estimations in Chapter 6.

Chapter 4

Estimation Considerations: Dealing with Sources of Potential Bias

OLS estimation of the labour market returns to human capital is often criticised for its susceptibility to two potential sources of estimation bias. The first relates to the omission of any explanatory variables that may be causally associated both with the dependent and the independent variables in a model while the second relates to the non-random selection of estimation samples. Although the literature on the implications of these sources of estimation bias is often divided on the most appropriate ways in which to deal with them, most studies emphasise that the failure to take cognisance of and control for any resultant endogeneity distorts the behavioural relationships that are of interest, biases parameter estimates, and undermines the fidelity of inferences that are drawn from estimation results.

The general nature of labour market returns to human capital estimations coupled with the specific features of the NIDS data outlined in Chapter 3 imply that estimation bias is also a potentially serious cause for concern in this study. The presence of any unobserved factors that influence not only educational attainment levels, but also the probability of being employed and on-the-job earnings capacity, for example, could lead to biased estimates of the labour market returns to education if they are not accounted for in some way. Similarly, if the patterns of non-missing observations on the NIDS numeracy and school quality variables, or the likelihood of being employed, are related to unobserved factors that also influence labour market earnings, unbiased estimation of the earnings returns to human capital may be hampered by the resulttant endogeneity.

In order to evaluate the scope for, assess the theoretical implications of, and formulate an appropriate strategy for dealing with potential biases in the estimations in Chapter 6, this chapter places the descriptive and formal definitions of omitted variable and sample selection bias in

the context of the primary objectives of this study - i.e. the estimation of employment and labour market earnings returns to educational attainment, school quality, and numeracy in South Africa. While the econometric literature proposes various theoretical approaches for dealing with these two sources of bias, the sample limitations in the NIDS data necessitate a strategy that achieves balance between accuracy and feasibility of implementation. Moreover, the chosen strategy may not be allowed to compromise examination of the explicit linkages between the three measures of human capital in the NIDS data and the two labour market outcomes of interest. In the discussion below, it is argued that the use of proxy variables in combination with the Heckman ML approach satisfies these prerequisites.

4.1 Omitted Variable Bias

In standard OLS regression, omitted variable bias ensues when a variable that influences both the outcome of the model and one or more of its explanatory variables is excluded from the estimation. In such an event, the underlying assumption that the conditional expected value of the OLS error term is equal to zero is violated and OLS estimates will no longer be unbiased.

In the context of Mincerean earnings functions, omitted variable bias most commonly arises due to the omission of a measure of ability from the regression estimation. In theory, failure to include an ability variable should bias OLS estimates of the returns to education upwards if higher ability is causally associated with both higher educational attainment and higher earnings (Keswell and Poswell, 2004, p. 846).[1] If such a relationship is unaccounted for, the regression coefficient(s) on the education variable(s) in earnings functions captures not only the marginal effect of incremental educational attainment on earnings, but also the marginal effect of higher innate ability levels. This is commonly known as *ability bias.*

Because of the limited scope of historically available micro level data on natural abilities and labour market outcomes in South Africa, very little research has been conducted to establish the potential extent of ability bias in South African education returns estimates. The most recent study to investigate this in South African earnings functions is that of Mariotti and Meinecke (2009), who estimate nonparametric bounds to the marginal earnings returns to education for Black males in South Africa. Controlling for sample selectivity and accounting for omitted ability in their estimates, the authors find that omission of ability from normal parametric earnings function estimations may bias the marginal returns to high school education upwards by between 3 and 5 percentage points (PP). In a previous study, Moll (1998, p. 275) finds that the inclusion of a measure of cognitive skill in the South African earnings function reduces the marginal return to education by between 6 and 12 percentage points.

[1] As discussed in Chapter 2, the notion that ability is positively correlated with educational attainment is theoretically supported by both the HCT and SH (Blackburn and Neumark, 1993, p. 522).

While ability bias is a particularly poignant example of omitted variable bias and features prominently in the labour market returns to education literature, it is important to note that, within the context of human capital returns estimations, other forms of omitted bias are similarly plausible. For example: if the quality of the education that individuals receive influences both their ability and propensity to accede to higher levels of education and their subsequent earnings capacity in the labour market, conventional Mincerian earnings functions may also be subject to what could be called *education quality bias* (hereafter referred to as quality bias).[2] Omission of a measure of education quality from the earnings estimation would again lead to endogeneity and cause the estimated returns to educational attainment to be biased upwards.

From the discussion above, it is not difficult to appreciate that the estimation of labour market returns to human capital in this paper may also be vulnerable to omitted variable bias in the guises of ability and quality bias. It is therefore appropriate to formally consider how these forms of bias are likely to influence estimates of the returns to human capital in theory and to consider strategies that could be used to compensate for their effects.

4.1.1 A Formal Definition of Omitted Variable Bias

The formal description of omitted variable bias can be facilitated by framing it within the context of OLS estimation of the labour market returns to education.[3] For a population comprising of N observations, a simple linear multivariate labour market outcome data generating process (DGP) can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector of observed labour market outcomes (in this instance, either the probability of employment or the log of earnings, depending on the outcome of interest); $\boldsymbol{X}$ is a $N \times K$ covariate matrix which includes a variable that measures the number of years of educational attainment, $\boldsymbol{x}_{educ}$, and $K-1$ other control variables (including a constant); $\boldsymbol{\beta}$ is the $K \times 1$ estimable parameter coefficient vector which relates the covariate matrix to the outcome variable, $\boldsymbol{y}$; and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of stochastic error terms. The OLS estimate of the parameter vector is found by minimising the sum of squared errors such that

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} = 0$$

[2] Case and Yogo (1999, p. 3) find evidence that school quality in South Africa does influence educational attainment levels.

[3] This is a context within which omitted variable bias is often framed. However, omitted variable bias is relevant in any context where variables that influence both outcomes and explanatory factors are excluded from estimation specifications. While this section abstracts from the formalisation of such examples, it is important to note that the formulations presented here are generalisable to other contexts.

Solving this equation and substituting for $\boldsymbol{y}$ yields the standard OLS result

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{OLS} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \end{aligned} \tag{4.2}$$

Equation 4.2 shows that $\hat{\boldsymbol{\beta}}_{OLS}$ will be an unbiased estimate of the true parameter vector $\boldsymbol{\beta}$ only if the conditional expected value of the stochastic error term is equal to zero. In other words, if $E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0$, the regression satisfies the assumption of exogeneity and OLS estimates will be consistent and unbiased. However, if the covariate matrix in equation 4.1 excludes variables that not only have a deterministic relationship with the outcome variable, but are also systematically related to one or more of the explanatory variables included in the covariate matrix, OLS estimation will produce biased results.

To explain why OLS estimates are biased when certain variables are omitted, consider the effects of cognitive ability and education quality on educational attainment and labour market outcomes. As discussed above, cognitive ability and school quality are not only likely to influence labour market outcomes, but also levels of educational attainment. Both ability and the quality of education are inherently difficult to measure and capture adequately in survey data and are therefore often excluded from earnings regression specifications. However, if they are not included in labour market returns to education estimations, their systematic relationships with the outcome variable are absorbed in the error term which, given the assumptions above, will by definition no longer be orthogonal to the covariate matrix. The appropriateness of the estimation specification used must therefore be evaluated against the theoretical full benchmark model which includes these variables. Letting $\boldsymbol{W}$ represent the $N \times 2$ covariate matrix containing the hypothetical ability and education quality variables[4], the true labour market outcomes DGP can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{4.3}$$

with

$$\boldsymbol{x}_{educ} = \boldsymbol{W}\boldsymbol{\delta} + \boldsymbol{\mu} \quad \text{for} \quad \boldsymbol{x}_{educ} \in \boldsymbol{X} \tag{4.4}$$

and

$$\boldsymbol{\gamma} \neq 0; \boldsymbol{\delta} \neq 0$$

Here, $\boldsymbol{\gamma}$ is a 2×1 parameter vector measuring the effect of unobserved innate ability and education quality on $\boldsymbol{y}$; $\boldsymbol{\delta}$ is a 2×1 parameter vector measuring the effect of innate ability and education quality on educational attainment; and $\boldsymbol{\mu}$ is the $n \times 1$ vector of stochastic error terms in the education function (Equation 4.4). If equation 4.3 is indeed the true model, exclusion of

[4] For J omitted variables, $\boldsymbol{W}$ would be a $N \times J$ covariate matrix

$\boldsymbol{W}$ when estimating equation 4.1 will produce

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{OLS} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{W}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{\gamma} \\ \therefore E\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right) &= \boldsymbol{\beta}+\boldsymbol{\gamma}(\boldsymbol{X}'\boldsymbol{X})^{-1}\Sigma_{\boldsymbol{XW}} \end{aligned} \tag{4.5}$$

Equation 4.5 shows that, unless $\boldsymbol{\gamma}$ is equal to zero or $\boldsymbol{X}$ is orthogonal to $\boldsymbol{W}$, which has been assumed not to be the case, OLS will produce biased estimates of $\boldsymbol{\beta}$. It may be possible to gauge the direction of the bias by inferring the signs on $\boldsymbol{\gamma}$ and $\Sigma_{\boldsymbol{XW}}$ from theory or *a priori* expectations. For example, if it can be presumed that ability is positively related to labour market outcomes (i.e. $\boldsymbol{\gamma} > 0$) and educational attainment (i.e. $Cov\left(\boldsymbol{x}_{educ}\boldsymbol{w}_{ability}\right) > 0$, omission of a measure of ability from the estimation of the labour market returns to education should cause the estimate of the rate of return to education, $\hat{\boldsymbol{\beta}}_{educ}$, to be biased upwards. However, even if the direction of the bias can be inferred accurately in this manner, the extent of the bias cannot be known without knowing $\boldsymbol{W}$. Consequently, it remains dubious to draw naive inferences from estimations where omitted variable bias is likely to be a matter for concern.

4.1.2 Attenuating Omitted Variable Bias through the use of Proxy Variables[5]

It should be evident from the discussion above that, by omitting measures of ability and school quality from earnings and employment returns functions, OLS estimation could potentially yield biased estimates of the labour market returns to education. The obvious solution would therefore be to incorporate these measures in estimations. However, the type of micro level survey data that allow for the analysis of labour market outcomes rarely include direct measures of ability or education quality. In the absence of these measures, researchers employ various estimation strategies which should theoretically diminish the extent of the omitted variable bias or even negate it altogether. These strategies include the use of instrumental variables (IV) and two-stage least squares (2SLS) estimation[6], fixed effects estimation using panel data[7], within-group difference estimation with data on monozygotic twins[8], or other nonparametric estimation techniques[9].

Before resorting to any of the approaches above, it is worth considering that omitted variable bias could aslo be mitigated by including variables in the estimation that serve as proxies for

[5] The notation and description of proxy variables in this section is based on Wooldridge (2002, 2009, pp. 63-67; pp. 284-288) and Greene (2002, pp. 87-88)

[6] See, for example, Chen and Hamori (2009).

[7] See, for example, Fertig and Schurer (2007) and Dolton and Silles (2008).

[8] See, for example,Ashenfelter and Krueger (1994); Ashenfelter and Rouse (1998); Behrman and Rosenzweig (1999); Arias *et al.* (2001); Bonjour *et al.* (2003) and Isacsson (2004).

[9] See, for example, Mariotti and Meinecke (2009).

unobserved covariates. Closely related to the unobserved variables, these proxies effectively become surrogate carriers of unobserved, but relevant information in the regression model. Continuing with the notation from Section 4.1.1, assume again that $\boldsymbol{W}$ is omitted because it cannot be directly observed. Furthermore, assume that for each omitted variable, $\boldsymbol{w}_j \in \boldsymbol{W}$, there is one proxy variable, $\boldsymbol{z}_j \in \boldsymbol{Z} \notin \boldsymbol{X}$, which is observed in the data.[10] The relationship between the omitted variables and the proxy variables can then be expressed as

$$\begin{gathered} \boldsymbol{w}_j = \boldsymbol{z}_j \boldsymbol{\alpha}_j + \boldsymbol{\nu}_j \quad \text{for} \quad j = 1, \ldots, J \\ \therefore \boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\nu} \end{gathered} \tag{4.6}$$

where $\boldsymbol{W}$ is a $N \times J$ matrix of unobserved variables; $\boldsymbol{Z}$ is the $N \times J$ matrix of proxy variables; $\boldsymbol{\alpha}$ is a $J \times J$ parameter matrix relating each of the proxy variables in $\boldsymbol{Z}$ to its corresponding omitted variable in $\boldsymbol{W}$; and $\boldsymbol{\nu}$ is a $N \times J$ matrix of stochastic error terms in the latent equation. Crucially, $\boldsymbol{Z}$ can only be a valid proxy for $\boldsymbol{W}$ if it is correlated with $\boldsymbol{W}$. In other words, in equation 4.6 it must hold that $\boldsymbol{\alpha}_j \neq 0 \; \forall \; j \in J$.

It can be shown that OLS estimation with proxy variables will yield consistent and unbiased estimates of $\boldsymbol{\beta}$ when two conditions are satisfied. First, the proxy variables must have no explanatory power in the main equation over and above their function as proxies for the omitted variables. In other words, if it were possible to include the previously unobserved omitted covariates in the regression, inclusion of the proxies would be completely redundant. Formally, this requires that $E(\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z}) = E(\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{W}) = 0$ such that $E(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z}) = E(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})$. Of course, since the omitted variables are not actually observed, this condition is not a testable constraint. However, Wooldridge (2009, p. 285) notes that the assumption is true (practically) by definition and that it is therefore not particularly controversial.

The second condition is more restrictive and requires that the covariates contained in $\boldsymbol{X}$ do not explain any of the residual variation in $\boldsymbol{W}$ once the proxy variables in $\boldsymbol{Z}$ have been controlled for. This requirement holds only if, in addition to the assumption that $E(\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z}) = 0$, it is assumed that $E(\boldsymbol{\nu}|\boldsymbol{X}, \boldsymbol{Z}) = 0$ such that $E(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Z}) = E(\boldsymbol{W}|\boldsymbol{Z}) = Z\boldsymbol{\alpha}$. Once again, this condition is not empirically testable since $\boldsymbol{W}$ is unobserved. If it does not hold, $\boldsymbol{Z}$ is an imperfect proxy for $\boldsymbol{W}$ and OLS will still produce biased estimates of the $\boldsymbol{\beta}$ parameter vector.[11] However, Bekker and Wansbeek (1996, p. 302) find that even when proxy variables violate the second assumption for a perfect proxy, they may still serve to mitigate the extent of omitted variable bias under OLS estimation. By implication, whenever omitted variable bias is a concern

[10] Lubotsky and Wittenberg (2006) generalize the use of proxy variables to cases where there are multiple proxies for each omitted variable. However, for the purposes of this study, the one-to-one proxy to omitted variable treatment is sufficient.

[11] Wooldridge (2002, p. 64) notes that there is some confusion in the literature about the appropriate use of the proxy variable classification and that authors need to be explicit about its definition where applicable. In this paper, a proxy variable refers to a variable which satisfies the redundancy condition, but which may or may not necessarily satisfy the second condition.

and proxy measures are available, it is preferable to use them, however imperfect they may be, rather than to exclude them from the estimation.[12]

The upshot of the proxy variable approach is that, in the context of returns to education estimation, it may be possible to diminish the extent of *ability* and *quality bias* by using proxy measures like cognitive assessment test scores (such as IQ, literacy, and/or numeracy test scores) and indicators of school performance which, while not necessarily commensurate to, are supposed to be reflective of cognitive ability and education quality. Of course, the validity of such an approach depends on the validity of the proxy variables that are used. The adequacy of the NIDS numeracy test score and school quality measure as proxies for ability and education quality have already been outlined implicitly in Sections 3.2.1 and 3.3.1. Including these two variables in estimations of the labour market returns to education thus serves a dual purpose. First, it allows for the estimation of returns to components of human capital that are not fully reflected by educational attainment levels. Second, it serves as an attempt to attenuate the bias in OLS estimates of the returns to educational attainment that results from the omission of unobserved ability and education quality measures.

Given that the NIDS numeracy and school quality scores are at best imperfect proxy measures of ability and education quality, it is expected that there may still be some edogeneity in the returns to education estimations. Moreover, from the results in Sections 3.2.2 and 3.3.2 it appears as though the pattern of censoring on these two variables is correlated with the omitted variables for which they are to stand proxy. Using them to account for omitted variable bias may therefore have the unintended consequence of introducing additional sample selection bias (Du Rand *et al.*, 2011, pp. 10 - 11).[13] In order to account for this possibility, it is necessary to consider the implications of selection bias, not only insofar as it relates to the selective pattern of non-missing observations of the NIDS numeracy and school quality variables, but also in terms of the selection of individuals into labour force participation and employment.

4.2 Selection Bias

In regression estimation, sample selection bias obtains as a consequence of failing to adequately control for systematic, unobserved differences between samples for which an outcome variable is observed and samples for which it is not (Heckman, 1979). Consider, for example, an earnings function estimation where the outcome variable, earnings, is only observed for individuals

[12] This is not meant to imply that the proxy variable approach is necessarily superior to any of the alternative approaches advocated for dealing with omitted variable bias. However, given the specific features of the NIDS data (as discussed in Chapter 3) it may be particularly suitable for the estimations in this study.

[13] Du Rand *et al.* (2011) employ an IV strategy in an attempt to account for selection on the NIDS numeracy and school quality variables when they are included as regressors in earnings estimations. However, their results show that this approach *"fails"* due to a lack of adequate instruments with sufficient numbers of non-missing observations in the NIDS data (Du Rand *et al.*, 2011, p. 16).

who, by virtue of being employed, receive labour market remuneration. The process of selection into employment and, by extension, selection into the earnings estimation is not arbitrary, but instead is systematically related to various observed and unobserved individual-specific characteristics, many of which not only influence the likelihood of employment, but also the earnings that individuals receive once they are employed. If the factors that affect both earnings and the probability of observing earnings cannot be accounted for due to unobservability, the explanatory variables in the earnings estimation will be correlated with the error term (Vella, 1998, p. 192). In other words, the failure to fully control for the endogenous nature of the selection process will cause OLS estimates of the earnings function to be biased.

While the potential for sample selection bias and the need to account for it is most commonly emphasised within the context of earnings function estimations, it is an issue which needs to be acknowledged whenever the censoring pattern on an outcome variable appears to be non-random. In this paper, for example, there are four potentially endogenous selection processes that may need to be accounted for: selection into the labour force, selection into employment, selection into the NIDS numeracy test module, and selection into the group of NIDS respondents who have school quality information available. Before continuing with a brief description of each of these selection processes, some clarification on the concept of selection is needed. In this paper, selection refers to any process whereby a particular outcome of interest is observed for only a sub-sample of a population, irrespective of how, why, by whom or through what the process is effected.[14] In other words, selection is not merely confined to scenarios where the observation of an outcome depends on individuals' behaviours or decisions.

In the estimation of the employment and earnings returns to education, school quality, and numeracy, it is necessary to take cognisance of four distinct cases of potentially endogenous selection mentioned above. First, in order to become employed, individuals need to participate in the labour force.[15] If the propensity to participate in the labour force is related to the unobserved correlates of employment probability, failure to account for selection into participation could lead to biased OLS estimates of the employment returns to human capital. The second selection process manifests in terms of selection into earnings as has already been outlined above. Third, as shown in Section 3.3.2, the sample of individuals for whom school quality data is available appears to differ systematically from those for whom it is not. Specifically, it seems as though younger individuals with higher levels of educational attainment and potentially higher levels of unobserved innate ability are more likely to have school quality data available. Since these and other potentially unobserved characteristics should also be related to the probability of em-

[14] The definition used here best serves the purposes of this study and does not necessarily correspond to other definitions of selection that are found in the literature.

[15] It is worth noting that the decision to participate in the labour force may itself be endogenous. Imagine, for example, a situation where an economically inactive student is offered a job by a close relative. The prospect of guaranteed employment may be a sufficient incentive for the student to start participating in the labour force and accept the job. In this case, the probability of employment explicitly influences the probability of labour force participation.

ployment and earnings capacity, selection into school quality data may need to be accounted for when using the school quality variable as a covariate in labour market returns estimations. Finally, following a similar rationale to the selection on school quality data, the individuals who participated in the NIDS numeracy test and therefore have numeracy scores seem to differ markedly from those who did not participate in the test. Again, the correlates of numeracy test participation are likely to be correlated with numeracy test performance, employability and earnings capacity. Selection into the numeracy test thus also needs to be accounted for in the labour market returns estimations.

4.2.1 A Formal Definition of Selection Bias[16]

To formally illustrate why OLS produces biased estimates in the presence of endogenous selection, consider again the example of the earnings function estimation described above. The outcomes of the estimation depend on two stages, each of which can be modelled using a separate equation:

$$1^{st}\ Stage\ Selection \begin{cases} \boldsymbol{z^*} = \boldsymbol{X_1}\boldsymbol{\alpha} + \boldsymbol{\nu} \\ z_i = 1 \quad if \quad z_i^* > 0 \\ z_i = 0 \quad if \quad z_i^* \leq 0 \end{cases} \tag{4.7}$$

$$2^{nd}\ Stage\ Outcome \begin{cases} \boldsymbol{y^*} = \boldsymbol{X_2}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ y_i = y_i^* \quad if \quad z_i = 1 \\ y_i = . \quad if \quad z_i = 0 \end{cases} \tag{4.8}$$

where z_i is a binary indicator that is equal to 1 if an individual's earnings is observed (i.e. if an individual is employed[17]) and equal to zero if not; y_i is observed earnings, z_i^* and y_i^* are the respective latent counterparts of the employment outcome in the first-stage equation and the earnings outcome in the second-stage equation; $\boldsymbol{X_1}$ is a matrix of observable variables that determine whether an individual is employed or not; $\boldsymbol{X_2}$ is a matrix of observable variables that determine the earnings of employed individuals; $\boldsymbol{\alpha}$ is the parameter vector that relates the covariate matrix, $\boldsymbol{X_1}$, to the latent employment outcome; $\boldsymbol{\beta}$ is the parameter vector that relates the covariate matrix, $\boldsymbol{X_2}$, to the latent earnings outcome; and $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$ are the stochastic error terms which capture the effects of any unobserved correlates of employment status and earnings that are not already included in the covariate matrices $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$.

As mentioned above, it is conceivable that many of the factors that influence employment status also influence earnings capacity. Consequently, it is likely that the covariate matrices, $\boldsymbol{X_1}$ and

[16] The notations and descriptions of sample selection bias in this section are based on Davidson and MacKinnon (2004, pp. 478-481), Cameron and Trivedi (2009, pp. 541-545), Puhani (2000, pp. 54-56), Greene (2002, pp. 782-784), Burger (2008, pp. 1-4), and Johnston and DiNardo (1996, pp. 447-449),

[17] For the sake of simplicity it is assumed that earnings are only observed when an individual is employed.

$\boldsymbol{X_2}$, will contain some of the same explanatory variables. However, it is not commonality between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ that leads to sample selection bias, but rather commonality between the the first- and second-stage error vectors, $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$. If any of the unobserved factors that explain employment (e.g. innate ability) also explain earnings, the first and second-stage error terms will be correlated. To show why this causes OLS estimates to be biased, equations (4.7) and (4.8) can be used to express the expected value of earnings conditional on the observation of earnings as

$$\begin{aligned} E\left(y_i^*|z_i=1\right) &= E\left(\boldsymbol{x_{2}}_i\boldsymbol{\beta}+\varepsilon_i|z_i=1\right) \\ &= \boldsymbol{x_{2}}_i\boldsymbol{\beta}+E\left(\varepsilon_i|z_i=1\right) \\ &= \boldsymbol{x_{2}}_i\boldsymbol{\beta}+E\left(\varepsilon_i|z_i^*>0\right) \\ &= \boldsymbol{x_{2}}_i\boldsymbol{\beta}+E\left(\varepsilon_i|\boldsymbol{x_{1}}_i\boldsymbol{\alpha}+\nu_i|>0\right) \\ E\left(y_i^*|z_i=1\right) &= \boldsymbol{x_{2}}_i\boldsymbol{\beta}+E\left(\varepsilon_i|\nu_i>-\boldsymbol{x_{1}}_i\boldsymbol{\alpha}\right) \end{aligned} \tag{4.9}$$

Equation 4.9 shows that whenever the second-stage error, $\boldsymbol{\varepsilon}$, is not orthogonal to the first-stage error, $\boldsymbol{\nu}$, OLS estimation of the outcome equation will yield biased parameter estimates.

4.2.2 Accounting for Selection Bias using the Heckman ML model[18]

Various strategies have been proposed to allow for the unbiased estimation of parameters when endogenous selection is a concern. Among the most commonly used approaches is the Heckman (1979) two-step sample selection correction procedure. Since this approach is not appropriate when working with complex survey data like NIDS 2008, this section proposes the use of the maximum likelihood (ML) version of the Heckman selection procedure to account for the four cases of selection outlined above (StataCorp, 2009*a*, p. 76).[19] However, before discussing this approach it is useful to consider the Heckman (1979) two-step estimator, not only because it facilitates the discussion of the ML model, but also because the two-step estimates are generally used as starting values for the estimation of the ML model.

Heckman (1979) observed that, under certain conditions, the problem of sample selection bias, as illustrated in equation 4.9, reduces to a case of omitted variable bias. Assuming that the

[18] The notations and descriptions of the Heckman (1979) two-step and ML sample selection correction models in this section are based on Cameron and Trivedi (2009, pp. 542), Wooldridge (2002, p. 566), StataCorp (2009*a*, p. 658), Nawata (1994, p. 34) and Puhani (2000, pp. 54-56).

[19] In the presence of sample weighting, stratification, clustering and/or not independently distributed data, the probabilistic interpretations of standard log-likelihood functions no longer hold. Instead, parameters need to be estimated by maximising the associated pseudo log-likelihoods which take the survey design into account. This implies that the Heckman (1979) two-step procedure, which entails estimation via OLS in the second-step, cannot produce consistent parameter estimates when working with complex survey data. It is therefore necessary to estimate the two stages of the Heckman (1979) model simultaneously via ML. (StataCorp, 2009*a*, p. 76)

first and second-stage error terms in equations 4.7 and 4.8 follow a bivariate normal distribution such that

$$\begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{\varepsilon} \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & \sigma_\varepsilon^2 \end{bmatrix} \right) \tag{4.10}$$

with

$$\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$$
$$\nu_i \sim N\left(0, 1\right)$$
$$cor\left(\varepsilon_i, \nu_i\right) = \rho$$

where N_2 denotes the bivariate normal distribution, the second-stage error term can be expressed as a function of the first-stage error:

$$\boldsymbol{\varepsilon} = \rho\sigma_\varepsilon \boldsymbol{\nu} + \boldsymbol{\mu} \tag{4.11}$$

with

$$cor\left(\nu_i, \mu_i\right) = 0$$

Substituting equation 4.11 into equation 4.9, the expected value of the outcome equation conditional on the observation of an outcome becomes

$$\begin{aligned} E\left(y_i^* | z_i = 1\right) &= \boldsymbol{x_{2i}}\boldsymbol{\beta} + E\left(\rho\sigma_\varepsilon\nu_i + \mu_i | \nu_i > -\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right) \\ &= \boldsymbol{x_{2i}}\boldsymbol{\beta} + \rho\sigma_\varepsilon E\left(\frac{\nu_i}{\sigma_\varepsilon} | \frac{\nu_i}{\sigma_\varepsilon} > -\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right) \\ &= \boldsymbol{x_{2i}}\boldsymbol{\beta} + \rho\sigma_\varepsilon E\left(\frac{\phi\left(-\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right)}{1 - \Phi\left(-\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right)}\right) \\ &= \boldsymbol{x_{2i}}\boldsymbol{\beta} + \rho\sigma_\varepsilon \frac{\phi\left(\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right)}{\Phi\left(\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right)} \end{aligned} \tag{4.12}$$

where ϕ is a standard normal density function and Φ is a cumulative density function. Equation 4.12 once again illustrates that OLS estimates will be biased if $\rho \neq 0$. Heckman (1979) argued that the selection bias in equation 4.12 is analogue to omitted variable bias and suggested that, given the assumption of bivariate normality and the computability of the first-stage selection equation outcomes, a variable can be constructed which, when included in the second-stage outcome specification, would compensate for the bias arising from endogenous selection (Johnston and DiNardo, 1996, p. 449). This variable, which is commonly referred to as the inverse mills ratio (IMR) and is generally denoted by the greek symbol λ can be expressed as

$$\lambda\left(-\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right) = \frac{\phi\left(-\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right)}{\Phi\left(-\boldsymbol{x_{1i}}\boldsymbol{\alpha}\right)} \tag{4.13}$$

Substituting the IMR into equation 4.12 yields

$$E\left(y_i^*|z_i=1\right)=\boldsymbol{x_{2i}\beta}+\rho\sigma_\varepsilon\lambda\left(-\boldsymbol{x_{1i}\alpha}\right)$$

The inclusion of the IMR in the outcome equation specification thus compensates for the bias term in equation 4.12. As the name suggests, the two-step approach involves two estimation steps. In the first step, the selection equation 4.7 is estimated using a probit model in order to obtain valid estimates of $\boldsymbol{\alpha}$.[20] In the second step, the predicted values from the selection estimation are used to construct an estimate of the IMR, $\hat{\lambda}$, which is then included as an additional covariate in the second-stage outcome estimation. Subsequent OLS estimation of the outcome equation should now produce unbiased estimates of $\boldsymbol{\beta}$ (Davidson and MacKinnon, 2004, p. 480). Equation 4.8 thus becomes

$$\boldsymbol{y^*}=\boldsymbol{X_2\beta}+\theta\lambda\left(-\boldsymbol{x_{1i}\alpha}\right)+\boldsymbol{\varepsilon} \tag{4.14}$$

While the parameter estimates produced with OLS estimation of equation 4.14 will be unbiased, endogenous selection implies that their standard errors will be subject to heteroscedasticity. Since censoring on the second-stage outcome variable, $\boldsymbol{y}$, depends on the values of the first-stage explanatory variables, the variance of the residual, $\boldsymbol{\varepsilon}$, can be shown to vary over observations with $\boldsymbol{x_{1i}}$:

$$Var\left(\varepsilon_i\right)=\sigma_\varepsilon^2-\left(\rho\sigma_\varepsilon\right)^2\left[\boldsymbol{x_{1i}\alpha}\cdot\lambda\left(\boldsymbol{x_{1i}\alpha}\right)+\lambda\left(\boldsymbol{x_{1i}\alpha}\right)^2\right] \tag{4.15}$$

Under normal OLS, robust standard errors could be used to account for heteroscedasticity. However, the inclusion of an estimate of the IMR, $\hat{\lambda}$, which is itself a function of an averaged estimate of the first-stage selection equation parameter vector, $\hat{\boldsymbol{\alpha}}$, in the second-stage equation introduces additional variance into the model which would not be adequately accounted for with robust standard errors (Johnston and DiNardo, 1996, p. 449). Puhani (2000, p. 55) therefore recommends using the White (1980) heteroscedasticity adjustment or bootstrapping methods of estimation in conjunction with the Heckman approach.

The tractability and relative computational simplicity of the Heckman two-step estimation procedure has made it one of the most frequently employed approaches for dealing with endogenous sample selection. However, the approach is not without its disadvantages and should not be used indiscriminately (Johnston and DiNardo, 1996, p. 449). First, the Heckman (1979) model is wholly dependent on the assumption of jointly normally distributed error terms in the first- and second-stage equations (Puhani, 2000, p. 58). Second, evidence from Monte Carlo simulations suggest that the approach may be much more sensitive to equation mis-specification and the presence of heteroskedasticity than other approaches, including IV regression (Johnston and

[20] The bivariate normality assumption of the Heckman (1979) approach requires that a probit model be used to estimate the selection equation.

DiNardo, 1996, p. 450).

Perhaps the greatest disadvantage of the Heckman (1979) approach relates to the need for valid exclusion restrictions to be present in the first-stage selection equation. Identification in the model arises from the non-linearity of the IMR in the second-stage equation. As explained above, the first- and second-stage covariate matrices, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, may include common explanatory variables. In principle, the model would be identified even if $\boldsymbol{X_1} = \boldsymbol{X_2}$ and all the explanatory variables in the selection equation appear in the outcome equation. However, given that the IMR is simply a non-linear function of the explanatory variables in the first-stage estimation, such an equivalence would imply a high degree of multicollinearity between the estimated IMR term and the covariates in the outcome equation, leading to inflated standard errors, insignificant t-statistics, and unreliable inference. Moreover, if the explanatory variables in the first-stage equation do not sufficiently explain the selection process, such that there is little variation in the predicted selection probabilities, the estimated IMR can be closely approximated by a linear function of $\boldsymbol{X_1}$, exacerbating the aforementioned implications for the second-stage estimation. Accurate estimation of the second-stage equation therefore requires both that the probability of being selected is adequately explained by the variation in the first-stage explanatory variables in $\boldsymbol{X_1}$, and that $\boldsymbol{X_1}$ contains at least one valid exclusion restriction – i.e. at least one explanatory variable that explains selection, but does not explain the second-stage outcome and therefore is not included in $\boldsymbol{X_2}$. (Burger, 2008, p. 5)

As stated above, the two-step is not appropriate when working with complex survey data such as NIDS 2008. However, the Heckman (1979) model can also be estimated via ML. While the ML version of the model is often advocated on the basis that it may produce more efficient parameter estimates than the two-step approach, it suffers from many of the same weaknesses as the two-step model, albeit not necessarily to the same extent. Monte Carlo simulations have shown that the ML model is more sensitive to violations of the assumption that the first and second-stage error terms are distributed bivariate normal with mean zero than the two-step model (Wooldridge, 2002, p. 556). By contrast, Nawata (1994, p. 33) finds that the ML model may outperform the two-step model whenever there is a high degree of multicollinearity between the estimated IMR and the covariates included in the second-stage outcome equation.

The log-likelihood function for the Heckman (1979) model can be seen as comprising of two parts, each corresponding to one of the two distinct outcomes in equation 4.7. Using Bayes Rule, the joint probability of observing $z = 1$ and, therefore, also observing the second-stage outcome variable, y, can be written as

$$\begin{aligned} P\left(\boldsymbol{y}, \boldsymbol{z} > 0 | \boldsymbol{X_2}, \boldsymbol{X_1}\right) &= f\left(\boldsymbol{y}\right) \cdot P\left(\boldsymbol{z} > 0 | \boldsymbol{y}, \boldsymbol{X_2}, \boldsymbol{X_1}\right) \\ &= f\left(\boldsymbol{\varepsilon}\right) \cdot P\left(\boldsymbol{z} > 0 | \boldsymbol{y}, \boldsymbol{X_2}, \boldsymbol{X_1}\right) \end{aligned} \tag{4.16}$$

which, under the assumption of bivariate normality as in equation 4.10, can be re-written as

$$\begin{aligned} P\left(\boldsymbol{y}, \boldsymbol{z} > 0 | \boldsymbol{X_2}, \boldsymbol{X_1}\right) &= \frac{1}{\sigma_1}\phi\left(\frac{\boldsymbol{y} - \boldsymbol{X_2\beta}}{\sigma_\varepsilon}\right) \int_{\boldsymbol{X_1\alpha}}^{\infty} f\left(\boldsymbol{\nu}|\boldsymbol{\varepsilon}\right) d\boldsymbol{\nu} \\ &= \frac{1}{1}\phi\left(\frac{\boldsymbol{y} - \boldsymbol{X_2\beta}}{\sigma_\varepsilon}\right) \int_{\boldsymbol{X_1\alpha}}^{\infty} \phi\left(\frac{\boldsymbol{\nu} - \left(\boldsymbol{y} - \boldsymbol{X_2\beta}\right)\rho/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right) d\boldsymbol{\nu} \\ &= \phi\left(\frac{\boldsymbol{y} - \boldsymbol{X_2\beta}}{\sigma_\varepsilon}\right) \left[1 - \Phi\left(\frac{-\boldsymbol{X_1\alpha} - \left(\boldsymbol{y} - \boldsymbol{X_2\beta}\right)\rho/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right)\right] \\ &= \phi\left(\frac{\boldsymbol{y} - \boldsymbol{X_2\beta}}{\sigma_\varepsilon}\right) \Phi\left(\frac{\boldsymbol{X_1\alpha} + \left(\boldsymbol{y} - \boldsymbol{X_2\beta}\right)\rho/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right) \end{aligned} \tag{4.17}$$

where equation 4.17 represents the contribution to the likelihood for individuals for whom the outcome variable in the second-stage outcome equation is observed. The probability of observing $z = 0$ is not conditional on y and is therefore simply equal to the marginal probability that $z^* \leq 0$:

$$P\left(\boldsymbol{z} \leq 0\right) = P\left(\boldsymbol{\nu} \leq -\boldsymbol{X_1\alpha}\right) = \Phi\left(-\boldsymbol{X_1\alpha}\right) = 1 - \Phi\left(\boldsymbol{X_1\alpha}\right) \tag{4.18}$$

Combining equations 4.17 and 4.18 and simplifying, the total log-likelihood for the Heckman ML model can be expressed as:

$$\begin{aligned} l\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon, \rho | \boldsymbol{X_1}, \boldsymbol{X_2}\right) &= \boldsymbol{z} \cdot \ln\phi\left\{\ln\Phi\left(\frac{\boldsymbol{X_1\alpha} + \left(\boldsymbol{y} - \boldsymbol{X_2\beta}\right)\rho/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right) + \frac{\boldsymbol{y} - \boldsymbol{X_2\beta}}{\sigma_\varepsilon}\right\} \\ &\quad + \left(1 - \boldsymbol{z}\right) \cdot \ln\left(1 - \Phi\left(\boldsymbol{X_1\alpha}\right)\right) \\ \therefore \quad l\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon, \rho | \boldsymbol{X_1}, \boldsymbol{X_2}\right) &= \boldsymbol{z} \cdot \left\{\ln\Phi\left(\frac{\boldsymbol{X_1\alpha} + \left(\boldsymbol{y} - \boldsymbol{X_2\beta}\right)\rho/\sigma_\varepsilon}{\sqrt{1-\rho^2}}\right)\right. \\ &\quad \left. -\frac{1}{2}\left(\frac{\boldsymbol{y} - \boldsymbol{X_2\beta}}{\sigma_\varepsilon}\right)^2 - \ln\left(\sqrt{2\pi\sigma_\varepsilon}\right)\right\} \\ &\quad + \left(1 - \boldsymbol{z}\right)\ln\left\{1 - \Phi\left(\boldsymbol{X_1\alpha}\right)\right\} \end{aligned} \tag{4.19}$$

Notwithstanding the aforementioned criticisms of the Heckman (1979) model, this paper endeavours to exploit the flexibility of the Heckman ML approach while taking cognisance of its various potential shortcomings. However, over and above dealing with conventional issues of selection into labour force participation and employment, the estimations that follow may also need to account for selection into the NIDS numeracy test module and school quality data. Whenever the *NIDS numeracy score* and/or *school quality score* are used as covariates in labour market returns estimations there is additional censoring on the outcome variable due to the limited number of non-missing observations on these explanatory variables. It is important to note that this does not necessarily imply that the estimation of the outcome equation will be subject

to additional sample selection bias (Wooldridge, 2002, p. 567). However, as explained above, there is reason to believe that the unobserved correlates of censoring on the *numeracy score* and *school quality score* variables in the NIDS data may overlap with the unobserved correlates of employment probability and earnings capacity. If this is the case, inclusion of these variables as covariates in employment and earnings functions estimations will again lead to endogeneity.

Wooldridge (2002, pp. 567-570) suggests that selection on explanatory variables can be accounted for using standard instrumentation techniques in combination with standard methods to account for the selection on the dependent variable. However, given the data constraints on potential instruments for numeracy and school quality in the NIDS data, such an approach is not feasible. In order to account for endogenous selection on explanatory variables, this study therefore again implements the Heckman ML procedure.[21] By implication, whenever an outcome equation is subject to both non-random selection on the dependent variable and non-random selection on an explanatory variable, it is necessary to perform two instances of the Heckman ML. In the first case, the model is used to correct only for non-random selection on the dependent variable, ignoring any selection on the explanatory variable of interest. Thereafter, an attempt is made to correct only for non-random selection on the explanatory variable, ignoring any other selection on the dependent variable.[22] This approach, while admittedly imperfect, may allow one to gauge the extent to which non-random selection on the dependent variable and non-random selection on the explanatory variable, respectively, influence the estimates of the labour market returns to educational attainment, school quality and numeracy in South Africa. As such, the implementation of this procedure necessitates, in addition to the standard estimation requirements outlined above, accurate modelling of the selection processes that apply to the observation of school quality data and participation in the NIDS numeracy test module.

[21] If, for example, an earnings function is estimated using the NIDS numeracy score as one of the covariates, there are two selection processes at play. Here, observation of the earnings outcome variable is not only conditional on being employed, but also on having taken the NIDS numeracy test. Since the inclusion of a censored explanatory variable limits the sample of observed earners in the same way as being employed does, it would be reasonable to account for it in the same manner.

[22] Following Mohanty (2001*a*,*b*), Krishnan (1990), and Wetzels and Zorlu (2003), attempts were made to simultaneously account for both non-random selection on the dependent variable and non-random selection on an explanatory variable by modelling two first-stage selection equations jointly using a seemingly-unrelated bivariate probit regression, estimating two IMRs, and including these in the second-stage OLS outcome equation. However, as mentioned above, such an approach is not appropriate when working with complex survey data. In addition, the implementation of this approach via ML in the statistical package Stata proved to be non-trivially complicated and was therefore abandoned.

Chapter 5

Descriptives: Human Capital Stocks and Labour Market Outcomes in South Africa

The preceding chapters have highlighted the concepts related to, and the theoretical justifications for, the positive association between labour market outcomes and different indicators of human capital, reviewed some of the empirical evidence on the labour market returns to education, numeracy and school quality as found in the literature, discussed the strengths and weaknesses of the NIDS 2008 dataset as basis for empirical inquiry, and formulated a strategy with which to deal with potential sources of estimation bias. The discussion now turns to the empirical analysis of the returns to human capital in the South African labour market. As precursor to the multivariate regression analysis in chapter 6, this chapter commences with a descriptive overview of the states of, and relationships between, labour market outcomes and human capital stocks in South Africa, as inferred from the NIDS dataset. This descriptive analysis not only contextualises the estimation of the labour market returns to education, school quality, and numeracy in South Africa, but may also provide important priors against which to evaluate the findings of those estimations.

Given South Africa's historical context, it is rational to expect that demographic differences related to race, age, gender and geographical location will be correlated with labour market outcomes in the country. While the observed differentials in these outcomes are often viewed as the consequence of persistent labour market discrimination, recent studies suggest that much of what is commonly characterised as discrimination may be attributable to human capital acquisition differentials which are in turn invariably linked to sociodemographic characteristics in South Africa (Du Rand *et al.*, 2011). As a point of departure, this chapter therefore begins with an overview of the sociodemographic correlates of human capital acquisition in South Africa

and the differences in human capital holdings between different sub-groups of the population. It is shown that the sociodemographic composition of the South African labour market and the inequalities in labour market outcomes, while perhaps rooted in the discriminatory labour market policies of the Apartheid era, now persist due to the underlying dimensions that characterise human capital differentials. Finally, a brief non-parametric description of the relationships between educational attainment, school quality, and numeracy and labour force participation, employment, and earnings is provided and assessed.

5.1 Education, School Quality and Numeracy in South Africa

Table 5.1 presents the within-group means and standard deviations for the years of educational attainment, numeracy scores, and school quality scores in the NIDS dataset for different groups in the South African population of working age. The estimated figures suggest stark differences in educational attainment levels between different race groups and age cohorts. While the average level of educational attainment for Whites is estimated at just over 12 years of schooling, the average level of attainment for Blacks is only 8.74 years - less than the minimum compulsory schooling level (9 years) required in accordance with the South African Schools Act of 1996 (Wegner *et al.*, 2008, p. 422). While the average attainment for Asians is approximately 10.5 years, Coloureds have only marginally higher average attainment than Blacks at 9.14 years.

The low average levels of educational attainment for the Black and Coloured race groups is largely driven by low average levels of attainment for older age cohorts, as shown in table 5.1. Average educational attainment levels among the non-White population have increased dramatically in the past three decades, owing to the expansion of access to basic education in South Africa. This is reflected in Figure 5.1 which shows the mean levels of educational attainment for different race groups in NIDS by year of birth. It is clear from the graph that even before the democratization of South Africa there had been a substantial increase in average attainment levels for Blacks, Coloureds, and Asians. These improvements suggest a trend towards convergence in educational attainment levels between race groups. The average attainment levels reported in table 5.1 are therefore likely to exaggerate the extent of the attainment differentials for younger cohorts.

Figure 5.2 adds further support to the notion that racial average educational attainment differentials have become smaller over time by illustrating the educational attainment distributions for Black and White individuals in the 15 to 30 and 31 to 65 age categories using kernel density estimates. The attainment distribution for Blacks under the age of 30 is shown to be narrower and to have a higher mean than the distribution for individuals over 30. Compared to the older

Table 5.1: Educational attainment, school quality, and numeracy score means and standard deviations in South Africa

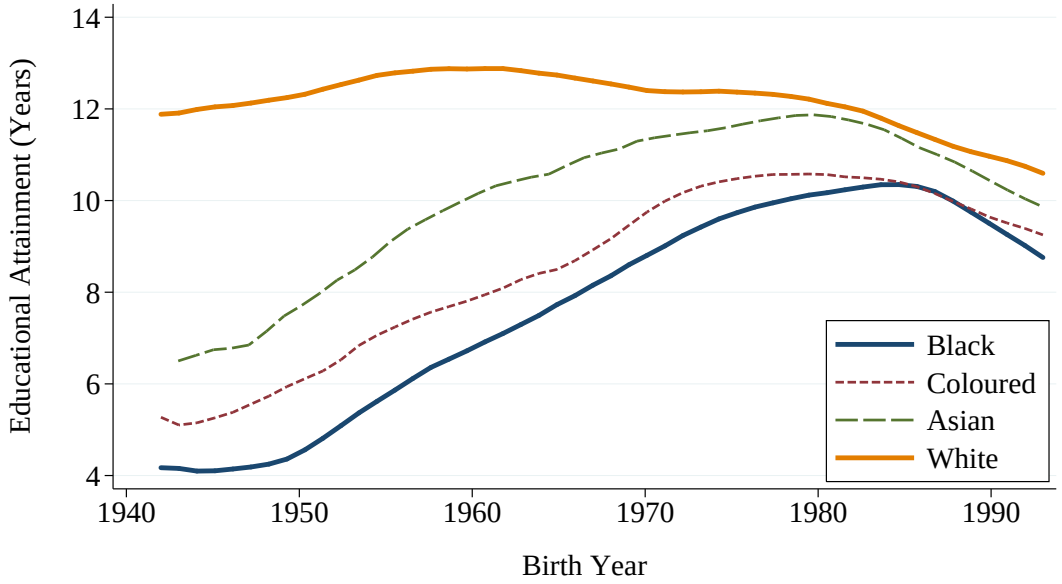
	Educational Attainment		**School Quality Score**		**Numeracy Score**	
Black	8.74	(3.72)	.06	(.07)	-.61	(1)
Coloured	9.14	(3.53)	.12	(.08)	-.33	(1.12)
Asian	10.44	(3.27)	.21	(.07)	-.51	(.97)
White	12.19	(2.34)	.24	(.08)	.18	(.86)
Male	9.21	(3.69)	.09	(.09)	-.51	(1)
Female	9.15	(3.75)	.09	(.09)	-.57	(1.04)
Urban	9.97	(3.3)	.1	(.1)	-.56	(1.04)
Rural	7.78	(4)	.07	(.07)	-.5	(.98)
15-19	9.29	(1.95)	.08	(.09)	-.51	(1.03)
20-24	10.51	(2.37)	.08	(.09)	-.57	(.93)
25-34	10.24	(3.08)	.08	(.09)	-.57	(.96)
35-44	9.02	(3.98)	.09	(.1)	-.57	(1.08)
45-54	7.72	(4.66)	.11	(.11)	-.53	(1.24)
55-65	6.24	(4.92)	.16	(.1)	-.26	(1.21)
Total	9.18	(3.72)	.09	(.09)	-.54	(1.02)

NOTES: Standard deviation estimates are reported in parentheses alongside the calculated means. The means and standard deviations for the *school quality* variable correspond to the estimates for the discritized standardized (mean: 0 stdev: 1) NIDS school quality measure whereas the means and standard deviations for the *numeracy* measure correspond to the estimates from the actual standardized (mean: 0 stdev: 1) NIDS numeracy score. Results are weighted.

age cohort, a greater proportion of young Blacks have attained tertiary levels of education (more than 12 years) while a smaller proportion failed to progress past primary education (7 years). In contrast to the findings for Blacks, Whites under the age of 30 not only have lower average levels of attainment than those over 30, but also have a far greater proportion of individuals with less than lower secondary (9 years) levels of attainment. As a result of the expansion of access to education in South Africa, the attainment distributions for Whites and Blacks in the younger age cohort therefore look far more alike than those for individuals in the older age cohort.

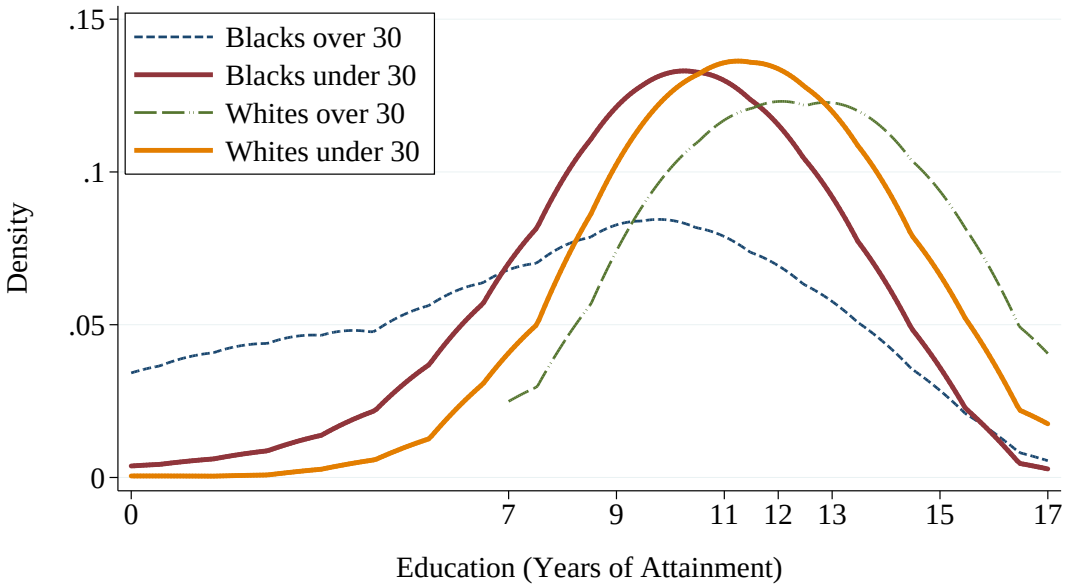
Similar to the racial educational attainment differentials discussed above, the large rural-urban average attainment differential reported in Table 5.1 is attributable to the large attainment differences between, in particular, Coloured and Black rural and urban individuals in the 31 to 65 age bracket. In reality, the South African rural-urban attainment divide has narrowed in the same manner as the Black-White divide illustrated in figure 5.2, though not quite to the same extent. Finally, in contrast to many other Sub-Saharan African countries, there does not appear to be a significant gender gap in average educational attainment levels in South Africa. Consistent with

Figure 5.1: Mean Educational Attainment by Race and Birth Year



NOTES: The local polynomial smoother used the Epanechnikov kernel with the default bandwidth and smoothing. The graph shows the estimated mean educational attainment level for each race group by birth year. Results are weighted.

Figure 5.2: Educational Attainment Distributions for different Black and White Age Cohorts



NOTES: The kernel densities were calculated using the Epanechnikov kernel with the default bandwidth. Solid lines denote estimates for individuals aged 15 to 30 whereas the dashed lines denote estimates for individuals aged 31 to 65. Results are weighted.

previous studies on educational attainment differentials, the NIDS data suggests that male and female average educational attainment levels are virtually identical, even amongst older South African cohorts (Anderson *et al.*, 2001, p. 41).

The diminishing gap between average educational attainment levels for different South African race groups is sometimes cited as a major achievement for the South African government in redressing the institutionalised discriminatory education policies and the racially delineated restriction of access to opportunities under the apartheid regime. However, the narrowing attainment gap belies the fact that the average quality of education received in South Africa remains highly variable both between race groups and also within the Black schooling system Case and Yogo (1999, pp. 4-6). The significant differences between the mean NIDS school quality scores presented in table 5.1 support the findings from previous studies investigating racial school quality differences in South Africa . The average secondary school quality score for White individuals is estimated to lie more than one standard deviation above the mean for Coloureds and about two standard deviations above the mean for Blacks. While average quality differentials are also observable between rural and urban areas and between different age cohorts, they are not as pronounced as the differentials between race groups.[1]

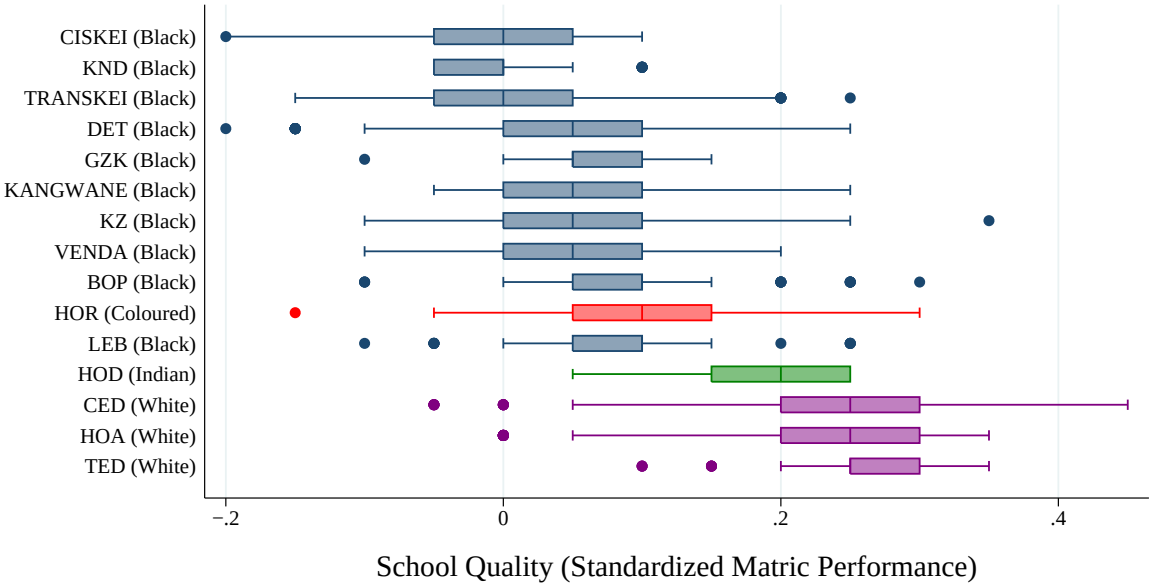
Persistent racial education quality differentials in South Africa are rooted in the racial and ethnic segregation of, and the inequitable distribution of resources between, the 18 former Apartheid-era education departments (Weber, 2002, p. 619).[2] The significant differences in the dispensations received by these departments invariably impacted on the quality of education that they were capable of providing to learners under their administrations. Figure 5.3 uses box plots to illustrate the variation in secondary school performance between and within each of the former education departments. The departments responsible for the administration of White learners (HOA, TED, CED) and Indian learners (HOD), vastly outperformed the formerly Coloured (HOR) and Black education administrations in terms of the quality of secondary school outputs as measured by the NIDS *school quality score* variable. It is similarly clear that there was substantial variation in the quality of schooling provided by the various departments that were previously collectively responsible for the education of Black individuals.

Despite the explicit de-racialisation of South Africa's formal education system in 1994, the

[1] It should be noted that these standard deviations are estimated for the discretized NIDS school quality measure opposed to the original standardized NIDS school quality score.

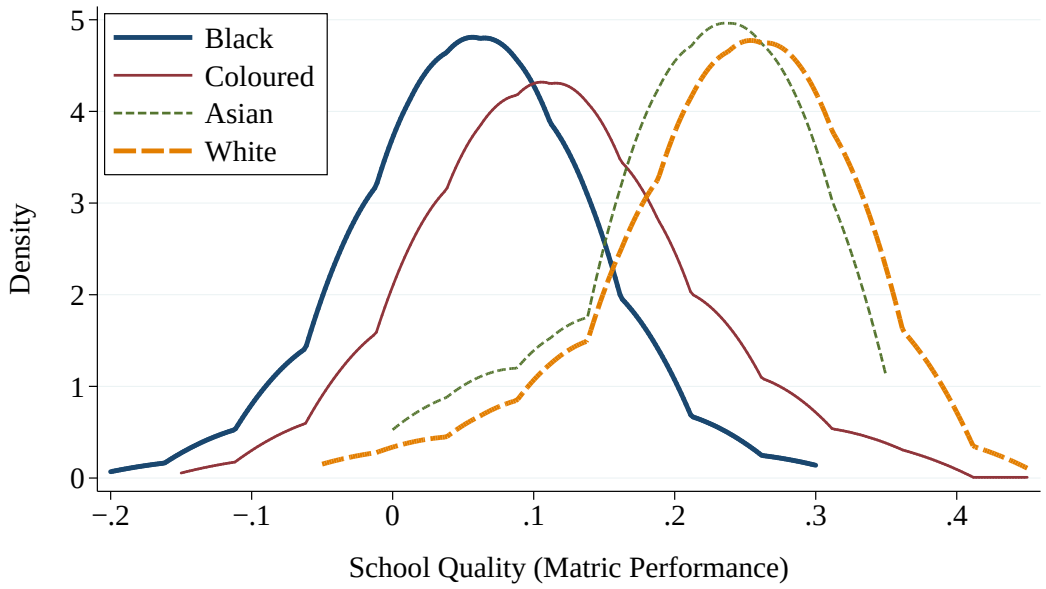
[2] Following the creation of the tricameral parliament in 1983, schools were administrated under 18 Departments of Education: Whites under the Department of Education and Culture: House of Assembly (HOA) and the administratively autonomous Transvaal (TED), Cape Province (CED), Orange Free State (OFS), and Natal (N) education departments; Coloureds under the Department of Education and Culture: House of Representatives (HOR); Indians under the Department of Education and Culture: (HOD); Blacks under the the Department of Education and Training (DET), the Bophuthatswana (BOP), Ciskei (CISKEI), Transkei (TRANSKEI), and Venda (VENDA) Homeland Education Departments, and the Departments of Education for the self-governing territories of Gazankulu (GZK), KaNgwane (KANGWANE), KwaNdebele (KND), KwaZulu (KZ), Lebowa (LEB), and QwaQw (QWAQWA) (Oosthuizen and Bhorat, 2006; Yamauchi, 2004, p. 16). (Weber, 2002, p. 620)

Figure 5.3: School Quality by Former Education Department



NOTES: Data for the Orange Free State (OFS), Natal (N), and QWAQWA education departments was not available in the NIDS data. Results are weighted.

Figure 5.4: School Quality Distributions by Race Group



NOTES: The kernel densities were calculated using the Epanechnikov kernel with a bandwidth of 0.05. Results are weighted.

persistence of socioeconomic inequalities - particularly relating to the limited spacial mobility of Blacks and Coloureds - implies that the racial composition of the majority of "formerly Black" and "formerly White" schools has remained virtually unchanged (Yamauchi, 2011, p. 148). While there has thus been an expansion in the access to education over the past decades, there has been little improvement in the expansion of access to quality education for non-White race groups (Motala, 2001, p. 66). The vast majority of Black, and to a lesser extent Coloured,

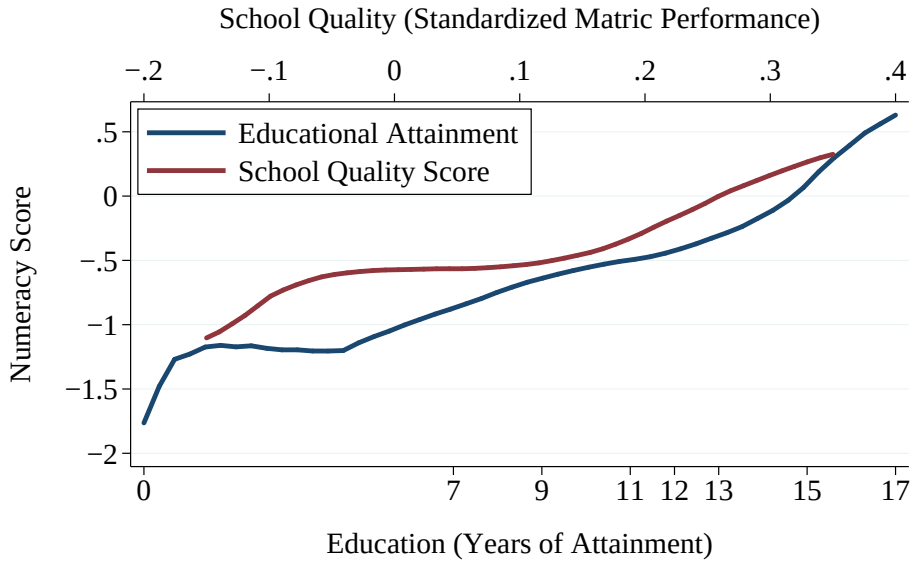
South Africans still have no other alternative but to attend historically disadvantaged schools.

The racial dimensions of the historical performance differentials between former education departments become clearer when considering the school quality distributions for different race groups in South Africa. The kernel density curves in Figure 5.4 are reflective of a dualistic education system where Whites and Indians have the means to attend functional schools that have retained the benefits of their historical advantage while the majority of Blacks and Coloureds lack the upward mobility required to escape the endemic dysfunctionally of a previously disadvantaged system. Moreover, bearing in mind that the NIDS *school quality score* is an aggregate measure of secondary school performance levels and is only available for a younger and more educated sub-sample of NIDS respondents, the kernel density estimates in figure 5.4 and table 5.1 offer an incomplete reflection of the extent of education quality differentials in South Africa and are likely to understate both the differences in the quality of education received by different race groups and the variation in the quality of education received by Blacks and Coloureds.

The extent of the school quality differentials between race groups in South Africa has important implications for the estimation of the labour market returns to human capital. If the quality of education is sufficiently distinct between race groups, race may actually serve as a proxy for education quality. It should then be possible to mitigate the magnitude of the unexplained "discriminatory" component of racial differentials in labour market outcomes by explicitly controlling for school quality in regression estimations. Using a simulation model to derive estimates of school quality from historical matric performance data and educational attainment levels, Burger and Van der Berg (2011) find evidence that the inclusion of a measure of school quality in earnings regressions accounts for a significant portion of what is otherwise perceived as discrimination in the South African labour market. Du Rand *et al.* (2011) reach similar conclusions using the school quality variable in the NIDS dataset.

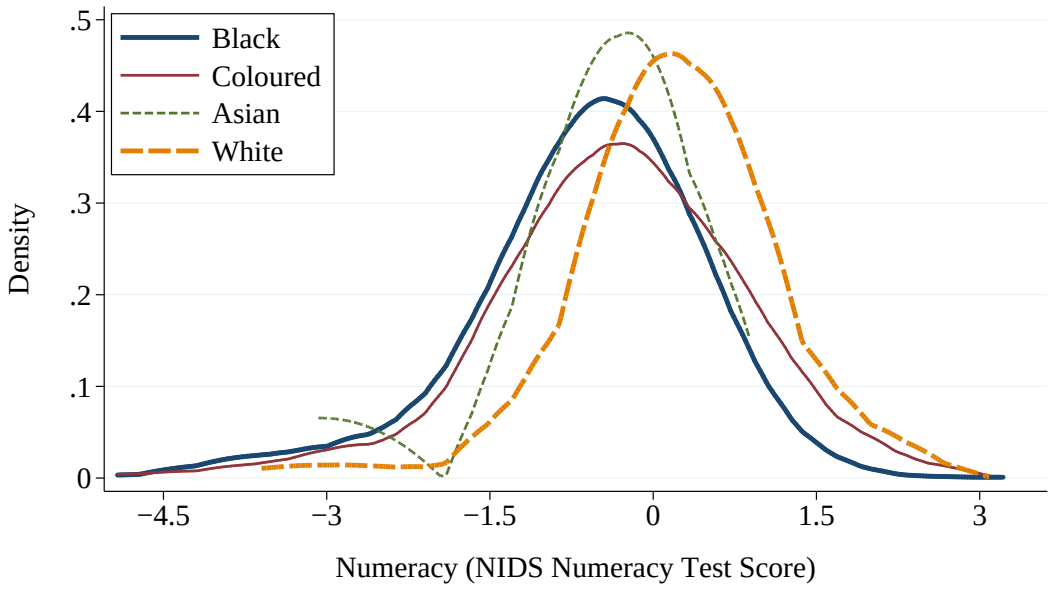
The educational attainment and school quality differentials discussed above also have important implications for numeracy levels in South Africa given that numeric competency is largely the product of innate ability, educational attainment en education quality. Figure 5.5 illustrates the positive associations between educational attainment and school quality and numeracy levels in South Africa. Given these relationships and the fact that there are no *a priori* reasons to assume different innate ability distributions for different population groups, it is not surprising that the differentials in numeracy levels presented in table 5.1 manifest along similar dimensions as the differentials in educational attainment and school quality levels in South Africa. However, from the data it appears as though the extent of the numeracy level differences between groups is not as pronounced as the differences in school quality. The mean White numeracy level is estimated to lie less than one standard deviation above the average for Blacks. There also appears to be less variation in average numeracy levels between the different age cohorts. As explained in Section 3.2.2, this result could obtain if individuals with higher levels of numeric ability were

Figure 5.5: Numeracy vs School Quality and Eduactional Attainment



NOTES: The local polynomial smoother used the Epanechnikov kernel with the default bandwidth and second-order smoothing. Results are weighted.

Figure 5.6: Numeracy Distributions by Race Group



NOTES: The kernel densities were calculated using the Epanechnikov kernel with a bandwidth of 0.5. Results are weighted.

also more likely to participate in the NIDS numeracy test module. In such an event, the NIDS numeracy test score distribution would most likely be left-censored and fail to accurately reflect the extent of variation in actually numeracy levels between different groups in South Africa.

Figure 5.6 illustrates the distribution of numeracy levels in South Africa for each race group. As expected, the numeracy distribution for Whites lies to the right of and is narrower than

the distributions for Blacks and Coloureds which are both relatively flat and, in the case of Blacks, virtually symmetrical around the overall mean numeracy score. It is important here to note again that, as in the case of the NIDS *school quality* score variable, it was mainly younger respondents with higher average levels of education who participated in the NIDS numeracy test and therefore have numeracy scores available. Consequently, it may reasonably be expected that the NIDS numeracy scores will be biased upwards and that the kernel density curves shown in figure 5.6 will therefore be excessively skewed to the right. Hence, while the Black and Coloured numeracy distributions presented in the figure are already quite flat, it is possible that the degree of within-group variation in numeracy levels for these two groups is significantly greater in reality.

5.2 Education, School Quality and Numeracy in the South African Labour Market

The discussion of educational attainment, school quality, and numeric competency differentials above provides an important background against which to assess the results of labour market returns to human capital stock estimations in South Africa. To gain an understanding of the context within which these returns are produced, it is furthermore useful to consider the sociodemographic composition of the South African labour market. Table 5.2 provides a breakdown of labour market status in South Africa by race, gender, geographical location, age, and educational attainment.

In 2008, South Africa's labour force[3] (LF) comprised of an estimated 18 395 044 individuals out of a working age population of approximately 30 507 529 people, resulting in a labour force participation (LFP) rate of 60.3%.[4] While rising average educational attainment levels and urbanisation among the Black population caused a dramatic rise in the LFP rate for Blacks in the first decade after Apartheid, table 5.2 shows that this group, despite constituting more than 75% of the total labour force, still had the lowest participation rate (58.48%) of all the race groups (Kingdon and Knight, 2007, p. 818). The same holds true for females who, despite experiencing markedly greater growth in LFP rates than males between 1993 and 2008, still had a lower LFP rate (61.23%) than men (71.93%) in 2008.

[3] The broad definition of the labour force is used throughout this paper.

[4] Leibbrandt *et al.* (2009*b*, p. 7) use NIDS 2008 to estimate the size of the South African labour force at 16 753 618 individuals with a labour force participation rate of approximately 55%. The reason for the difference between their figures and those reported here relates to the fact that they appear to have excluded data on NIDS proxy respondents from their calculations. However, the figures presented here are consistent with those in Statistics South Africa's 2008 4rth quarter statistical release of the Quarterly Labour Force Survey (Statistics South Africa, 2009, p. v).

LFP rates vary significantly by geographical location, age, and educational attainment levels. As is to be expected, a far greater percentage of individuals in urban areas (44.66%) participate compared to those living in rural areas (27.37%). There also appears to be a non-linear relationship between the LFP rate and age. The propensity to participate initially increases as individuals get older, but then starts to fall again after the age of about 40. Leibbrandt *et al.* (2009*b*, p. 7) provide evidence that the introduction of maximum age limits for each school grade in accordance with the South African Schools Act of 1996 was one of the primary reasons for the significant rise in LFP rate among individuals below the age of 30 after 1997. As a result of the growth in participation, 15 to 30 year-olds represented approximately 38% of South Africa's LF in 2008.

Given that higher levels of educational attainment are generally associated with higher probabilities of being employed, it is not surprising that LFP rates are found to be increasing in attainment. While participation rates are already high for individuals with lower and upper secondary schooling, table 5.2 shows a notable jump in the LFP rate after the completion of Matric - the level of attainment which many perceive to be a minimum requirement for successful entry into the job market. However, the unemployment rate for Matriculants[5] (30.23%) is only marginally lower than South Africa's overall unemployment rate of 31.06%.

The jump in labour force participation at the Matric level coupled with the high unemployment rates among participating Matriculants reflects three important aspects related to the South African labour market. First, when considered in isolation, the majority decision to participate in the labour market rather than to continue with tertiary study is a partial indication of the extent to which the financial costs associated with obtaining a tertiary qualification in South Africa make post-secondary education inaccessible to a large portion of the population. Second, the fact that the average educational attainment level of the South African labour force has risen steadily over the past two decades coupled with the widespread belief that the quality of education in South Africa has declined, implies that the labour market value of the Matric certificate has been eroded due to qualification inflation (Oosthuizen and Bhorat, 2004; Banerjee *et al.*, 2006; UMALUSI, 2005, p. 4). Third, the high rate of participation among Matriculants despite the high Matric unemployment rate suggests that many individuals overestimate the labour market value of the Matric certificate (Von Fintel and Black, 2007, p. 6).

While the 13.86% unemployment rate for individuals who have completed some form of tertiary education is considerably lower than the national rate, it remains paradoxically high given the existence of significant skills shortages in the South African labour market (Dias and Posel, 2007, p. 4). Oosthuizen and Bhorat (2004, p. 4), Akoojee *et al.* (2008, p. 254) and others have argued that the rising unemployment rates among young, tertiary-educated labour force participants is the result of structural changes in the labour market which have caused a mis-

[5] Individuals who have completed Matric.

Table 5.2: Labour Market Status and Sociodemographics in South Africa

	Row Percentages			*Column Percentages*			
	NEA	**Unemp**	**Empl**	**NEA**	**Unemp**	**Empl**	**Total**
Black	36.17	22.31	41.52	83.54	84.93	71.3	78.25
Coloured	26.75	18.95	54.3	6.99	8.16	10.55	8.85
Asian	24.26	11.12	64.63	1.99	1.5	3.94	2.78
White	25.03	10.97	64	7.48	5.41	14.22	10.13
Male	28.07	16.09	55.84	38.22	36.08	56.41	46.08
Female	38.78	24.36	36.87	61.78	63.92	43.59	53.92
Urban	27.37	20.61	52.03	50.57	62.72	71.35	62.55
Rural	44.66	20.45	34.89	49.43	37.28	28.65	37.45
15 to 19	79.51	11.2	9.29	39.72	9.21	3.44	16.91
20 to 24	32.93	32.79	34.28	14.65	24.02	11.31	15.05
25 to 34	14.81	27.99	57.2	11.56	35.97	33.11	26.4
35 to 44	14.48	20.63	64.9	7.71	18.1	25.65	18.03
45 to 54	26.5	14.54	58.96	10.69	9.66	17.65	13.65
55 to 65	53.27	6.28	40.45	15.68	3.05	8.84	9.96
No Education	47.4	14.71	37.89	9.85	5.04	5.86	7.04
Primary	38.3	19.55	42.15	21.06	17.71	17.26	18.64
Lower Secondary	47.06	21.95	31	24.05	18.49	11.8	17.33
Upper Secondary	39.17	22.8	38.02	30.11	28.89	21.77	26.06
Matric	19.75	24.26	55.99	11.36	23	23.99	19.5
Diploma	8.92	16.09	74.99	2.15	6.38	13.44	8.16
Bachelors	9.87	7.8	82.33	.34	0.45	2.14	1.18
Postgrad	17.66	0.49	81.85	1.09	0.05	3.75	2.08

NOTES: Row percentages denote the percentage of individuals within each row category who are respectively economically inactive (NEA), unemployed (Unemp), or Employed (Empl) whereas column percentages reflect the racial, gender, geographical, age, and educational composition of each column category (NEA, Unemp, Empl, Total). The broad definition of the labour force is used such that the unemployed also include discouraged work seekers. The estimation samples include data on NIDS proxy respondents. Results are weighted.

alignment between the skills that graduates have to offer and the skills that employers demand. The effects of this skills mismatch are exacerbated by the severe heterogeneity in the quality of education in South Africa, which serves to undermine the fidelity of education credentials as signals of potential labour market productivity (Mlatsheni and Rospabe, 2002, pp. 20-21).

Collectively, the factors outlined above imply that employment and unemployment rates in South Africa, as shown in table 5.2, vary along the same sociodemographic dimensions and in broadly the same direction, as educational attainment, school quality, and numeracy. Thus, unemployment rates are highest for younger, female, Black individuals in rural areas with low levels of educational attainment and lowest for White, middle-aged males in urban areas with tertiary qualifications.

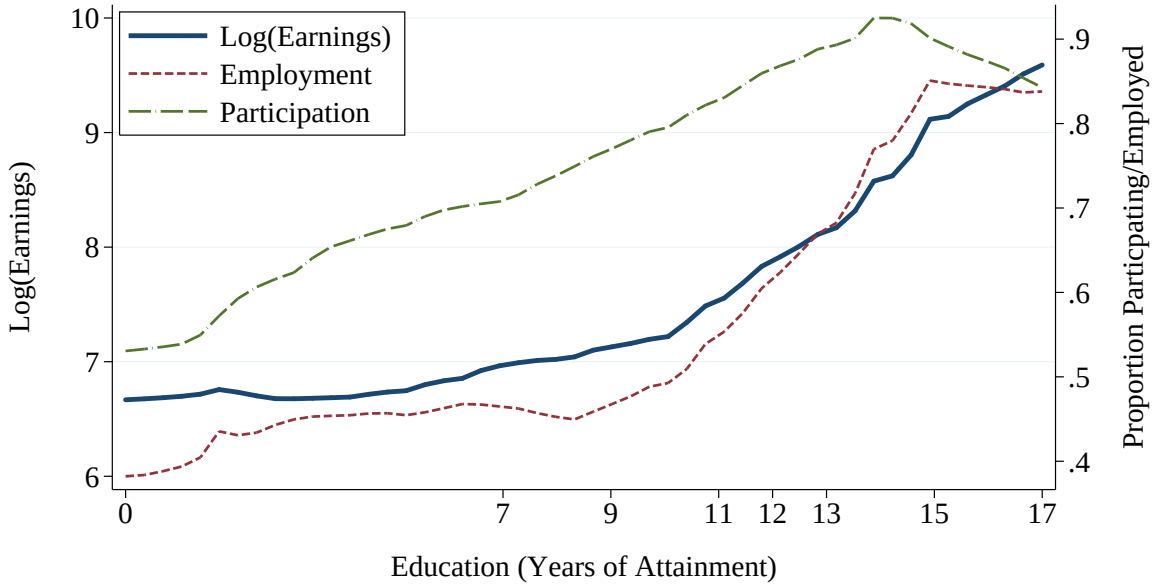
5.2.1 Employment and Earnings

The sociodemographic composition of the South African labour market, while partly rooted in the discriminatory labour market policies of the Apartheid-era, is to a large extent a function of the variation in human capital stocks across different sub-groups in the population. In other words, the observed differentials in labour market outcomes between characteristically distinct groups of South Africans are symptomatic of underlying differentials in their respective human capital stocks (as discussed in Section 5.1). To gauge the extent and structure of the associations between these human capital stocks and labour market outcomes in South Africa, various non-parametric graphical representations of the relationships between educational attainment, school quality, numeracy and labour force participation, employment, and earnings are given below.

Figure 5.7 shows that educational attainment has a relatively concave association with labour force participation and a convex relationship with employment and earnings in South Africa. This latter convexity is a common finding in the literature on average earnings returns to education in South Africa and is often hypothesised to be a consequence of the oversupply of unskilled labour in conjunction with the shortage of highly-educated, skilled labour in the economy (Keswell and Poswell, 2002, p. 20). However, convexity in the structure of the earnings and employment returns to educational attainment could also be explained by differentials in the quality of education in South Africa. If educational attainment is positively related to education quality, such that individuals towards the upper end of the educational distribution are, on average, also those who have attended better quality schools, then better-educated individuals should receive labour market premiums not only because of their higher attainment levels, but also because of the superior quality of the education they have received.[6] This would imply

[6] As explained in Section 2.2.2, this holds true in theory irrespective of whether education functions primarily as a signalling or a productivity augmentation device.

Figure 5.7: Labour Force Participation, Employment and Earnings by Educational Attainment



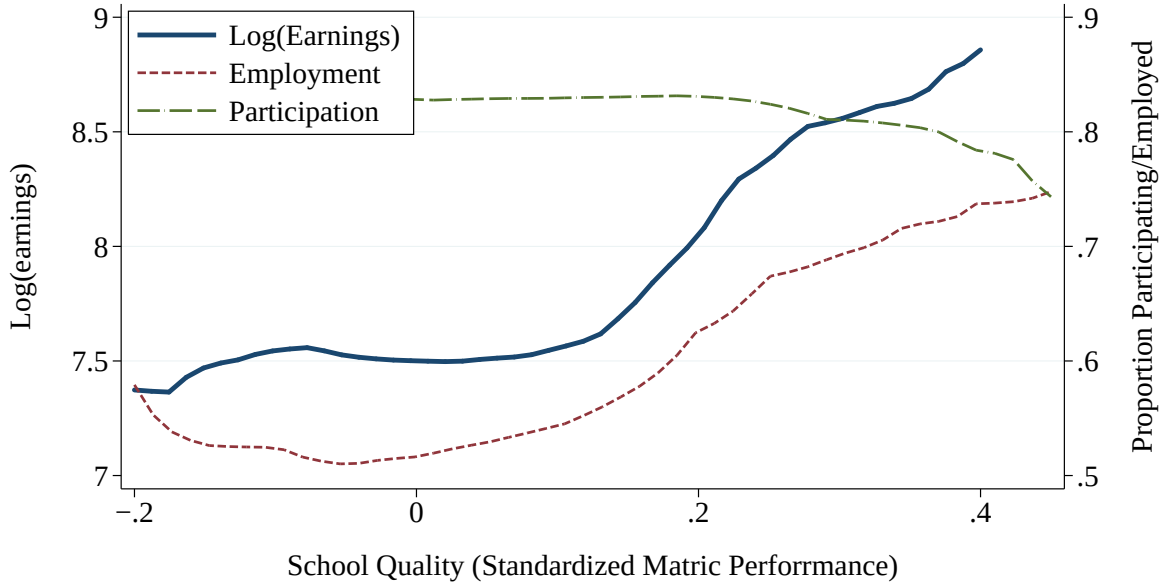
NOTES: Curves are drawn with zero-degree local polynomial smoothers that use the Epanechnikov kernel with the default bandwidth The estimation sample excludes individuals who indicated that they were enrolled in some form of formal eduction at the time of the survey. Results are weighted.

that convexity in returns could be observed even if the returns structures for low-quality-low-attainment individuals and high-quality-high-attainment individuals were both actually concave (Du Rand *et al.*, 2011, p. 8).

The relationships between school quality, labour force participation, employment and earnings that are illustrated in figure 5.8 are broadly similar to those shown in figure 5.7. Labour force participation appears to be concave in school quality while the employment and earnings structures are convex. When comparing the magnitudes of the changes in labour market earnings to changes in educational attainment and school quality through visual inspection of figures 5.8 and 5.7, it would appear as though the changes in average earnings and employment proportions are greater over the educational attainment distibution than over the school quality distribution. However, it must again be noted that the NIDS *school quality score* is, at best, only reflective of secondary school quality in South Africa and that, given the systematic processes underlying the observation of school quality data in the NIDS dataset as discussed in Section 3.3.2, it is likely that the lower tail of the NIDS *school quality score* distribution may be biased upwards. These issues imply that the school quality distribution may in reality be much wider than what is suggested in figure 5.8. It follows that the scope of the changes in labour market outcomes seen in figure 5.8 may understate the changes that would have been observed had the full South African school quality distribution been available.

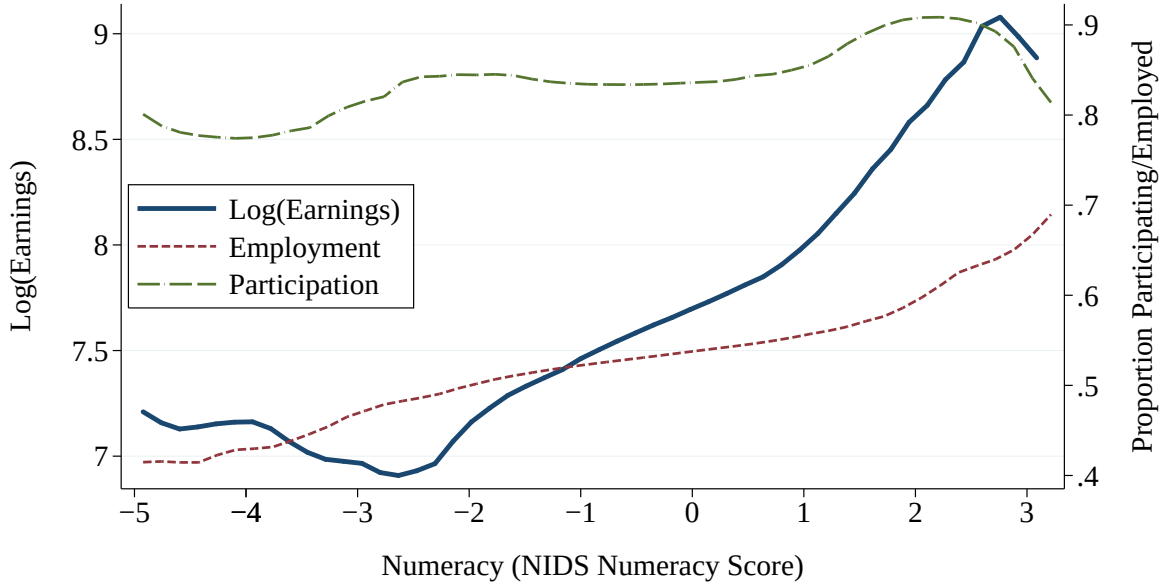
Figure 5.9 relates numeracy levels to labour market outcomes in South Africa. Here, the convexity between numeracy and earnings is even more pronounced than in figures 5.7 and 5.8. Following a similar rationale to the one above, it is again likely to be the case that the entire

Figure 5.8: Labour Force Participation, Employment and Earnings by School Quality



NOTES: Curves are drawn with zero-degree local polynomial smoothers that use the Epanechnikov kernel with the default bandwidth The estimation sample excludes individuals who indicated that they were enrolled in some form of formal eduction at the time of the survey. Results are weighted.

Figure 5.9: Labour Force Participation, Employment and Earnings by Numeracy



NOTES: Curves are drawn with zero-degree local polynomial smoothers that use the Epanechnikov kernel with the default bandwidth. The estimation sample excludes individuals who indicated that they were enrolled in some form of formal eduction at the time of the survey. Results are weighted.

NIDS numeracy score distribution is an upward-biased estimate of the real South African numeracy distribution. The earnings returns to numeracy at lower levels of the distribution in 5.9 probably overstate the average impact that low levels of numeracy has on labour market earnings. However, the return patterns in the figure do confirm that higher numeracy levels, whether truly reflective of greater innate ability or productivity, are associated with higher levels of labour market earnings.

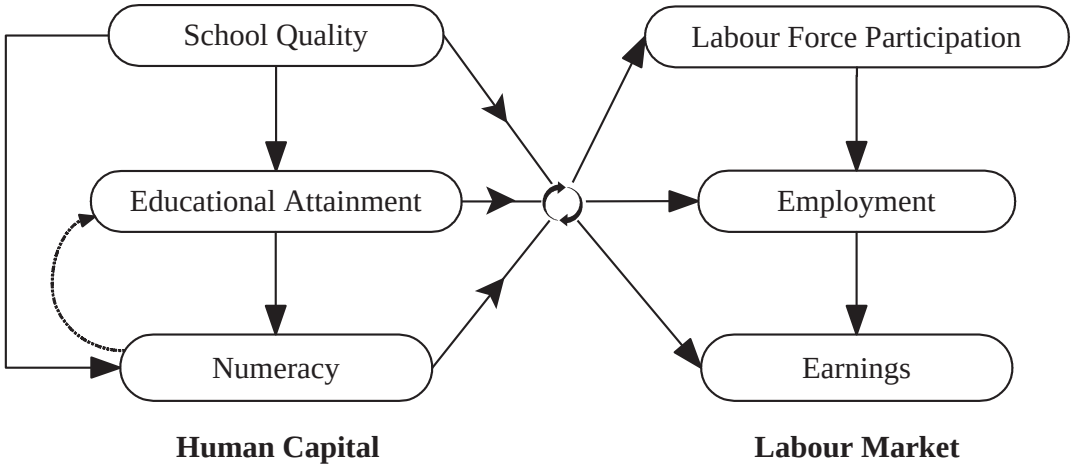
Chapter 6

Estimation: Labour Market returns to Education, Numeracy and School Quality in South Africa

The discussion thus far has highlighted several theoretical channels through which educational attainment, school quality, and numeracy are expected to influence labour market outcomes. The preceding descriptive analysis also illustrates some of the empirical manifestations of these relationships as observed for the South African labour market. To gain a deeper understanding of the true extent to which different components of human capital generate labour market returns in South Africa, it is necessary to explicitly account for the complex linkages that govern these relationships. Figure 6.1 provides a conceptual illustration of the causal relationships between the three human capital measures considered in this study and three primary labour market outcomes in South Africa.

As explained above, there are reasons to expect that school quality may have a causal influence on both educational attainment and numeracy. School quality is also likely to have a second-round indirect effect on numeracy through the causal relationship between educational attainment and numeracy. Insofar as numeracy is related to cognitive ability, it may in turn influence educational attainment. As a result, there may exist some self-reinforcing bi-directional causality between numeracy and educational attainment. The complex interdependencies that govern these relationships ultimately produce dynamic human capital stocks which can generate various labour market returns. However, the portion of these returns that accrue to each of the constituent components of a human capital stock cannot be observed directly. Yet, to understand which aspects of human capital are worthwhile for private or public investment, it

Figure 6.1: Conceptual Human Capital and Labour Market Linkages



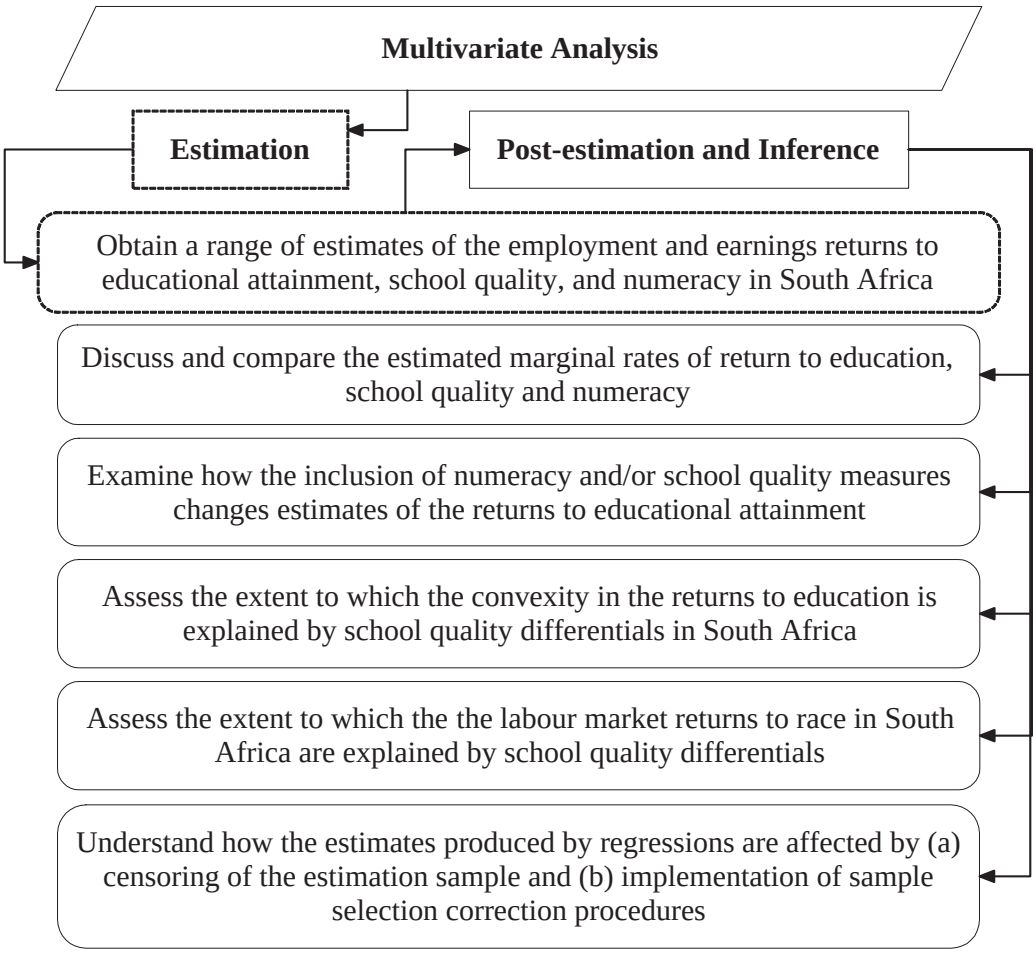
is necessary to know what their relative contributions are to those labour market outcomes that are observed. The aim of this chapter is therefore to conduct and discuss the results from the multivariate analysis of the NIDS dataset in an attempt to disentangle the effects that educational attainment, school quality and numeracy have on employment and earnings outcomes in the South African labour market.

6.1 Methodology

The multivariate analysis presented in this chapter can conceptually be separated into an estimation component and a post-estimation analysis and inference component. As shown in Figure 6.2, the former part enables the latter and, as such, should be conducted with care. The primary objective in the estimation analysis is to produce a range of estimates of the employment and earnings returns to educational attainment, school quality and numeracy in the South African labour market. In a latent sense, the goal is to obtain unbiased estimates of the labour market returns to these three measures of human capital. However, given the imperfections of the available data and the inherent complexity governing the theoretical identification of the parameters of interest, fully unbiased estimates are most likely unattainable. The goal therefore becomes to reduce, as opposed to eliminate, the bias to such a extent that there is a non-trivial increase in the fidelity of the estimates obtained. To this end, it would be prudent to produce not just one set of estimates, but a range of estimates as this should allow one to gauge how sensitive the results are to the model specifications and the approaches used.

The estimation results obtained in the first part of the analysis provides the empirical basis required to pursue five substantive objectives in the post-estimation analysis and inference. These objectives pertain to issues which either feature prominently in the political discourse surround-

Figure 6.2: Objectives of the Multivariate Analysis



ing the state of (and the policies related to) the South African labour market and education system or are common topics in the academic literature on the estimation of labour market returns to human capital and, more generally, the unbiased estimation of parameters in the presence of endogenous selection. Specifically, the objectives entail: (1) evaluating and comparing the relative importance of educational attainment, school quality and numeracy for South African labour market outcomes based on the estimates of the marginal returns to the respective human capital measures; (2) assessing the extent to which the omission of numeracy and school quality measures could bias standard OLS estimates of the returns to education in South Africa by examining the coefficient changes that occur when the NIDS numeracy or school quality variables are included in estimations; (3) examining the hypothesis that school quality differentials could provide a valid explanation for the strongly convex structure of the returns to schooling in South Africa; (4) assessing the extent to which the residual component of racial inequalities in labour market outcomes is attributable to racial differentials in the quality of education received; and (5) understanding how the labour market returns to human capital estimates are affected by the censoring of estimation samples and the implementation of procedures to correct for sample selectivity.

To best achieve the stated objectives, the estimation analysis is structured using the bottom-up estimation strategy illustrated in Figure 6.3. This strategy involves three sequential estimation stages. In the first stage, the employment and earnings returns to educational attainment, school quality and numeracy are respectively estimated via ML using probit models and via OLS using linear models, without attempting to control for any form of sample selection. Each model in this first stage is estimated using various alternative specifications in order to illustrate the effects that the inclusion or exclusion of certain explanatory variables and the censoring of the estimation samples have on the regression estimates.[1] The objective for the first stage in the strategy is to establish the baseline regression specifications to be used in the final estimation stage and to produce what will hereafter be referred to as "uncorrected" estimates of the labour market returns to educational attainment, school quality, and numeracy in South Africa. Here, "uncorrected" specifically refers to the fact that no attempt has been made to explicitly control for any endogeneity that may arise due to non-random sample selection.

In the second stage, the processes that determine whether individuals are included in final-stage estimation samples are modelled using probits and estimated via ML. As outlined in Section 4.2.2, the four selection processes that need to be modelled are: labour force participation (LFP) as a prerequisite for attaining employment; employment as a prerequisite for having non-zero earnings; being matched to a school as a prerequisite for being assigned a NIDS school quality score; and participation in the NIDS numeracy test module as a prerequisite for having a numeracy score.

The third and final stage in the estimation strategy is to re-estimate the baseline regressions from stage one, but to do so while attempting to account for the sample selection processes estimated in stage two. Initially, it is only selection on the outcome variable of interest - employment or earnings - that is accounted for. Thereafter, an attempt is made to account for any endogeneity resulting from the non-random observability of the explanatory variable of interest, namely either school quality or numeracy. The ML versions of the *Heckit* and *Heckprob* procedures, adjusted for weighting, stratification and clustering, are implemented to account for endogenous sample selection in these estimations.

The third stage in the estimation strategy produces various "corrected" estimates of the South African labour market returns to educational attainment, school quality, and numeracy which can be compared to the estimates produced in the first estimation stage in order to gauge the extent of the bias in the uncorrected returns estimates. The methodological approach followed is structured with the implicit aim of erring on the side of caution whenever inferences are drawn from the estimation results. The discussion of the important caveats to the estimation results and inferences drawn are deferred to the concluding Chapter of this study.

[1] The limited extent of overlap between the non-missing observations for the numeracy and school quality score variables precludes using them together as regressors in a model. Where applicable, models therefore include either the numeracy measure or the school quality measure.

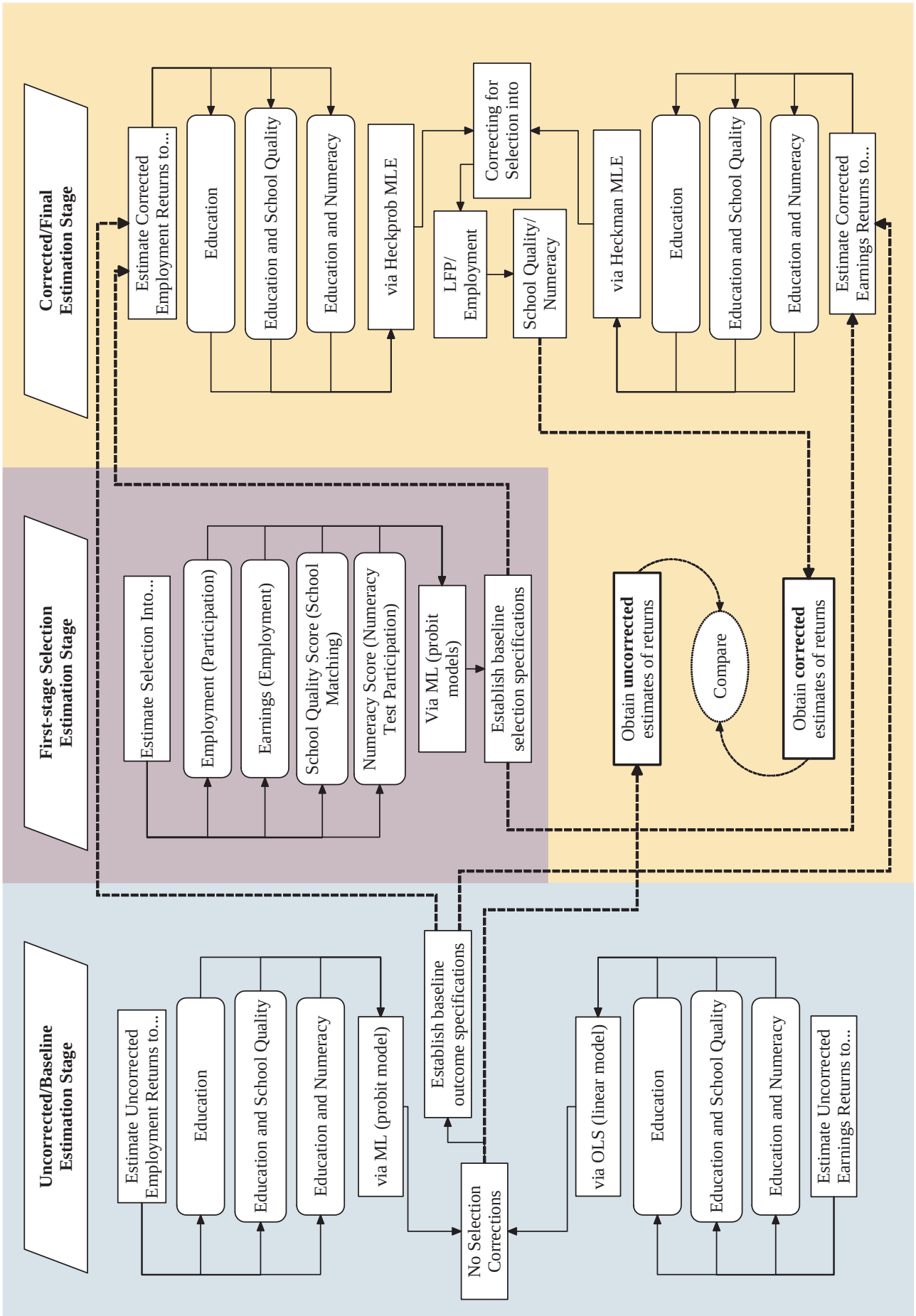


Figure 6.3: Estimation Methodology

6.2 Uncorrected Estimation[2]

6.2.1 Uncorrected Employment Returns

Table A.1 summarises the estimation results for the employment returns to educational attainment. A cumulatively more complete model is built from column (1) to (4). When controlling only for age (and age-squared), gender, geographical location, marital status, household head status, and race in model (1), Whites are seen to have a distinct advantage over other race groups in terms of procuring employment. In fact, Whites are predicted to have a 12.89% higher probability of being employed, on average, than Blacks, all other things being equal.[3]

To control for the impact of physical well-being on the probability of being employed, the regression specification in column 2 includes several dummy variables that capture individual's perceptions of their own health status and a dummy variable indicating whether they have a disability.[4] To also account for the relationship between emotional well-being and the probability of being employed, a control for individual emotional well-being is included. This measure is a composite index of a number of highly correlated indicator variables pertaining to respondent's self-reported emotional states, including the frequency with which individuals felt "unusually bothered", depressed, restless, lonely, lethargic, struggled to focus, or considered it an effort to complete tasks. The measures were collapsed into a single index using multiple correspondence analysis (MCA), with the loadings on the index suggesting that it is reflective of emotional well-being.

The results in Table A.1 show that the probability of being employed increases with improvements in perceived physical well-being and is negatively related to being disabled. Employment probability is also positively associated with emotional well-being.[5] However, the inclusion of these controls only marginally impacts on the magnitude and significance of the positive coefficient on the White racial dummy variable. It is only once educational attainment is controlled for in column (3) that the unexplained probability of employment premium for Whites becomes

[2] The analysis presented in this chapter employs various sociodemographic control variables in the regression estimations. It is acknowledged that many of the estimation results pertaining to these control variables may be interesting in their own right and warrant further discussion or investigation. However, the focus in this study falls squarely on the relationships between three specific human capital measures and labour market outcomes in South Africa. Unless the estimates on the control variables relate directly to the objectives outlined in the previous section, they are therefore either discussed only cursorily or not at all.

[3] As per the calculated average marginal effect on the White dummy variable in regression (1).

[4] The NIDS survey also asked a number of questions related to diagnosed medical conditions. However, in general the response rates on these questions were very low.

[5] It is important to qualify that one cannot infer the direction of causality purely from this result. Since NIDS respondents' labour market statuses were fixed at the time when they were asked to report on their emotional states, it is plausible that the result merely suggests that being employed raises emotional well-being on average. Alternatively, there could be unobserved factors driving the positive association.

statistically insignificant. This is an important result, as it suggests that Whites have no additional advantage over Coloureds, Indians, or Blacks in terms of the likelihood of procuring employment once inter-racial differentials in health, emotional well-being and educational attainment have been taken into account.[6]

Regressions (3) and (4) in Table A.1 account for educational attainment through a series of dummy variables that correspond to different levels of attainment in South Africa. This structure allows one to gauge the differences in the returns to education at different points in the attainment distribution. The results from both regressions show that pre-Matric levels of educational attainment do not generate any statistically significant employment returns. However, the estimated average marginal effects show that upon completion of matric the rate of return to education increases up to graduate attainment levels, decreasing only marginally for postgraduates.[7] Not surprisingly, individuals who are enrolled in education are found to be more than 20% less likely, on average, to be employed than their out-of-education counterparts.

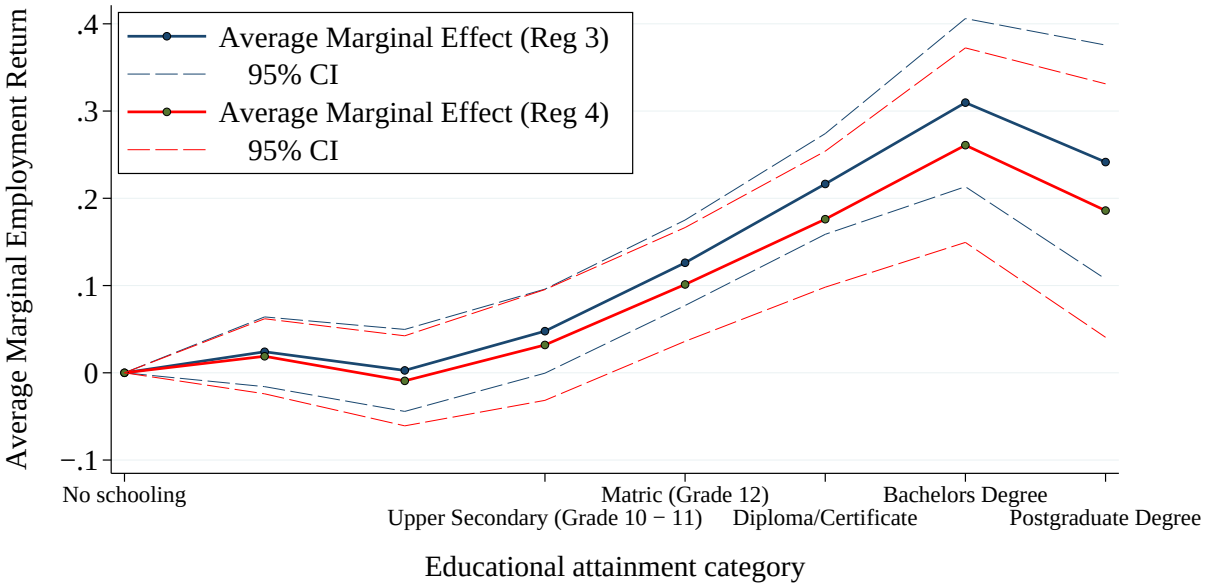
In an attempt to further reduce the potential bias in the returns to education estimates, model (4) includes two categorical variables that are indicative of the extent to which individuals consider themselves computer literate. Given the extent of modernisation in the South African labour market, it may be expected that computer literate individuals would find it easier to procure employment (Gustafsson, 2011, p. 33). The results in column (4) offer some support for this, suggesting that, on average, basic computer literacy may raise the probability of being employed by about 8.5% while advanced computer literacy raises the likelihood of employment by a further 8%, *ceteris paribus*. NIDS respondents were also asked to rate their own English reading and writing proficiencies. English being the *lingua franca* in the South African labour market, it is conceivable that individuals who are able to communicate effectively in English may find it easier to procure employment than those who cannot. To test this hypothesis, the English reading and writing scores were summed to create an "English competency" score. Column (4) shows that there does not appear to be any statistically significant employment returns to this measure of English competency once other human capital variables have been controlled for. Despite this fact, the inclusion of the computer literacy dummy and the English competency variable in regression (4) decreases the magnitudes of the estimated employment returns to education. This is also reflected in Figure 6.4 which shows that the average marginal employment returns to educational attainment estimates for model (4) lie below the estimates for model (3).[8]

[6] This result is robust to the exclusion of the "currently enrolled" dummy variable.

[7] Hypothesis testing shows that $H_0 : \bar{ROR}_{Matric} = \bar{ROR}_{Diploma}$ can be rejected at 1% significance and $H_0 : \bar{ROR}_{Diploma} = \bar{ROR}_{Degree}$ can be rejected at 10% significance. However, the hypothesis that the average rate of return for postgraduates is not statistically different from those for individuals with Matric, diploma or degree qualifications cannot be rejected at below-10% levels of significance.

[8] While the two average marginal return curves in the figure lie within one another's 95% confidence intervals, cross-equation hypothesis tests indicate that the estimates differences for the employment rates of return for individuals with post-Matric qualifications between the two equations are statistically significant at below-10% levels of significance.

Figure 6.4: Estimated Uncorrected Average Marginal Employment Returns to Educational Attainment



NOTES: Average marginal effects represent the discrete average percentage change in the probability of being employed relative to the base category (no schooling), evaluated over the observed values of the other regression covariates in the estimation sample: $\frac{\partial Pr(Employment)}{(Educaton_j - Education_{base})} = (\Phi(X\beta)|Educ_j) - (\Phi(X\beta)|Educ_{base}) \ \forall j \neq base$. The estimates reported here were calculated using the estimation results from regression (3) and (4) in Table A.1. The dashed lines represent the respective 95% confidence intervals.

The graph also reflects the increasing rate of return to educational attainment and, as such, is broadly consistent with the non-parametric estimates shown in Figure 5.7.

Model (4) in table A.1 controls for a fairly comprehensive selection of controls, the exclusion of which may otherwise have contaminated the estimated returns to education coefficients. The model fit is also reasonably good.[9] As such, this model represents the baseline specification for estimating the uncorrected employment returns to educational attainment. For the purpose of comparison, this specification is again used in column (1) of Table A.2 which extends the analysis to consider the impact of school quality on the probability of employment. It is important to note that the reference category for the educational attainment dummies in the regressions in Table A.2 is "completed primary education" whereas "no schooling" is used as the reference category for the regressions in Table A.1. This is the case because school quality data can only be observed for those individuals who actually went to school. It follows that regressions (2), (3), and (4) in Table A.2, all of which constrain the estimation sample to include only observations with school quality data available, do not include any information on individuals with no educational attainment.

[9] The sensitivity (proportion of correctly predicted "successes" for a given cutoff), specificity (proportion of correctly predicted "failures" for a given cutoff) and the area under the receiver operating characteristic (ROC) curve (measuring the predictive ability of the model at all cutoffs relative to uniformly distributed random variables at the chosen cutoff) suggest that the model does a fair to good job of accurately predicting successes and failures in the sample.

Column (2) in Table A.2 shows the results of the baseline model when estimated only for the sample of individuals who have school quality data available (hereafter referred to as the school quality sample). The model statistics show that this reduces the estimation sample size from 13 504 to 4 573 observations. There are also other changes in the estimated coefficients that offer support to the notion that the school quality sample may be characteristically distinct from the remainder of the working-age group in unobserved ways. The estimated return to post-graduate education in model (2), for example, is higher and of far greater statistical significance than the estimate for model (1). The lower levels of statistical significance on the estimated coefficients for Matric, diploma, and degree attainment levels also show that the standard errors on these estimates are inflated in the much smaller school quality sample. A plausible explanation for this result is that the lack of variation in educational attainment levels within the school quality sample, particularly at the lower end of the attainment distribution, coupled with the small size of the estimation sample leads to less precise estimates of the employment returns to education.

The school quality score variable in column (3) is not statistically significant and its inclusion does not appear to significantly change the coefficient estimates on any of the variables vis-à-vis the results in column (2). This result could obtain if the dummy variables controlling for race already sufficiently capture the impacts of school quality variation on the probability of employment.[10] However, even when the race dummies are omitted, as shown in column (4), the coefficient on the school quality variable remains statistically insignificant. From this result one may tentatively conclude that school quality, in the way that it is measured here and the sample within which it is captured, does not influence the probability of being employed once other variables have been controlled for.

Table A.3 presents the results from the estimations of the uncorrected employment returns to educational attainment and numeracy. Similar to the results presented in Table A.2, column (1) of Table A.3 displays the coefficient estimate for the baseline employment returns to education model when estimated on the full sample, column (2) shows the results from the baseline regression estimated only for individuals who have numeracy scores (hereafter referred to as the numeracy sample), model (3) adds the numeracy score as an explanatory variable, and model (4) omits the educational attainment dummies while retaining the numeracy score measure. Comparing the estimates from model (2) to those from model (1), it is clear that the numeracy sample differs substantially from the rest of the working-age sample. While the magnitudes of the returns to education coefficients are larger for the numeracy sample, none of the estimates are statistically significant.[11] In part, this could again be explained by the substantial reduction in the estimation sample size (from 13 504 to 3 395 observations) and the higher mean of and less variation in the educational attainment levels in the numeracy sample.

[10] As explained earlier, the mean NIDS school quality measure differs substantially by race group.

[11] The large and significant coefficient on the Asian race dummy is an artifact that arises from the fact that 14 of the 18 Asians included in the baseline numeracy sample are employed.

The results in columns (3) and (4) show that numeracy is positively related to the probability of being employed at the 10% and 5% levels of significance, respectively. However, even when the dummy variables controlling for educational attainment levels are excluded as in model (4), the estimated average marginal effects indicate that a one standard deviation increase in numeracy score is, on average, only associated with a 2.5% increase in the probability of employment, *ceteris paribus*.

The baseline regression results for the numeracy and school quality samples, when compared to those for the full sample, highlight an important concern with the estimation of employment returns to numeracy and school quality using the NIDS data. When a sample selection process produces a sufficiently homogeneous estimation sample, it may not be possible to estimate certain parameters of interest with any reasonable level of precision. Note that this result could hold even if the sample selection process itself was completely defined in terms of observables. In other words, the failure to obtain statistically significant estimates of the employment returns to educational attainment for the numeracy and school quality samples is not necessarily the result of endogenous selection into the respective samples, but may simply accrue to a lack of data variation in the estimation sample.[12]

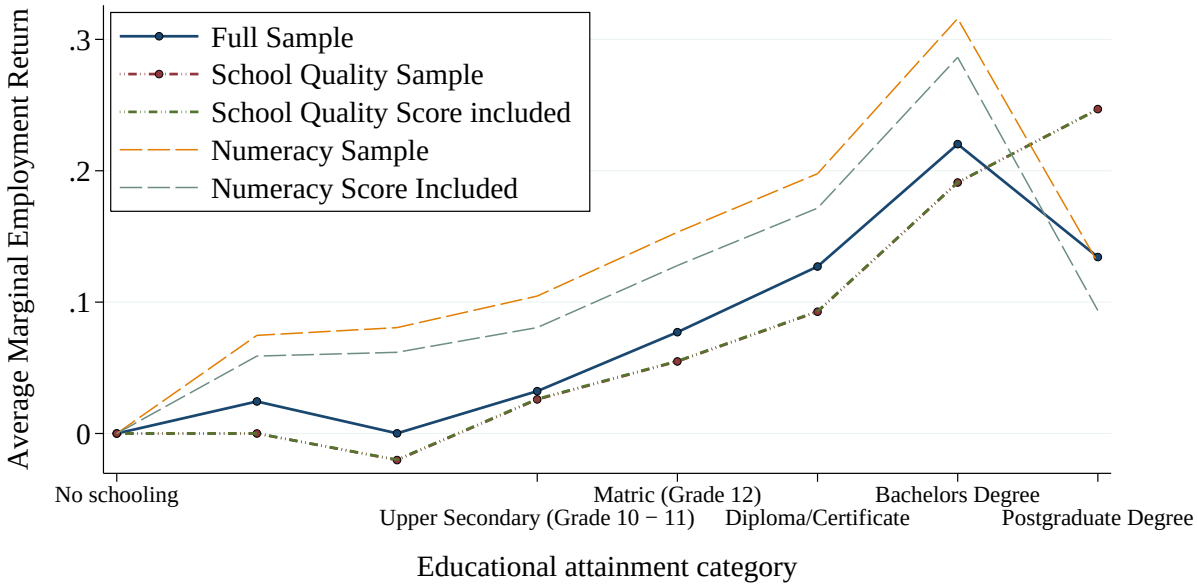
Acknowledging these concerns, the average marginal returns to educational attainment estimates from the various permutations of the baseline uncorrected employment returns estimation are illustrated graphically in Figure 6.5. The graph shows that the returns structure for the school quality sample is slightly flatter than that for the full sample. However, the marginal effects estimates for regression (3) in Table A.2 (which includes school quality as an explanatory variable) are virtually indistinguishable from those for regression (2). The current set of results therefore do not offer any support for the hypothesis that school quality differentials could account for the observed convexity in the employment returns to education structure in South Africa. This issue will be revisited when estimating the corrected employment returns to educational attainment in Section 6.4.

6.2.2 Uncorrected Earnings Returns

Having considered the set of uncorrected estimates for the employment returns to educational attainment, school quality, and numeracy in South Africa, the analysis now turns to the estimation of the uncorrected earnings returns to these human capital measures. Table A.4 presents the results for various permutations of the uncorrected earnings returns model. Most of the

[12] A good example of this is the reduced variation in educational attainment levels for the numeracy and school quality samples relative to the full sample. The standard deviation of the years of educational attainment in the full baseline estimation sample is 3.75 years with a mean of 9.11 years. By contrast, the standard deviation of educational attainment in the school quality and numeracy baseline estimation samples are 2.00 with a mean of 10.90 years and 2.46 with a mean of 10.28 years respectively.

Figure 6.5: Estimated Uncorrected Average Marginal Employment Returns to Educational Attainment for Different Estimation Samples



NOTES: Average marginal effects represent the discrete average percentage change in the probability of being employed relative to the base category (no schooling or completed primary), evaluated over the observed values of the other regression covariates in the estimation sample: $\frac{\partial Pr(Employment)}{(Educaton_j - Education_{base})} = (\Phi(X\beta)|Educ_j) - (\Phi(X\beta)|Educ_{base}) \ \forall j \neq base$. The estimates were calculated using the estimation results from regression (4) in Table A.1, (2) and (3) in Table A.2, and (2) and (3) in Table A.3.

covariates used in the employment regressions in the previous section are also used here in the earnings regressions. The estimates from the specification in column (1) (where only age (and age-squared), household head status, gender, marital status, geographical location and race are controlled for) not only show large marginal returns to education but also indicate that the earnings returns to education structure in South Africa is strongly convex. Moving from column (1) to column (5), the upward bias in the point estimates of the returns to educational attainment is shown to decrease as additional explanatory variables are added to the model.

Model (2) in Table A.4 adds three job-related variables to the estimation to control for the effects of job-specific characteristics on earnings. These include self-employment and casual employment dummy variables which are both interacted with a set of variables that capture the skill-level associated with an individual's occupation.[13] The inclusion of these variables shows that much of the variation in earnings which was initially explained by educational attainment in column (1) is actually attributable to variation in the types of the jobs that individuals have and the nature of their employment.

[13] Using a similar classification to the one employed by Posel and Casale (2011, p. 449) (who also use the NIDS 2008 dataset), individuals employed in elementary occupations are defined as unskilled; clerks, service workers, shop and market salespeople, skilled agricultural or fishery workers, craft and related trades workers, and plant and machinery operators or assemblers are defined as semi-skilled; and legislators, senior officials, managers, professionals, technicians and associate professionals are defined as skilled.

Similar to the findings for employment returns in Table A.1, emotional well-being and computer literacy are found to have significant positive effects on earnings. In contrast to the employment returns results though, English competency is found to be significant and positively related to earnings - even when controlling for other human capital variables. The marginal effects estimates suggest that individuals who are very competent at reading and writing English may earn up to a 30% earnings premium, on average, over those who read and write English very poorly, *ceteris paribus*.[14]

The results in column (5) show that being part of a labour union is associated with a 42.45% earnings premium, on average, in South Africa, all else held constant.[15] The strong positive relationship between union membership and earnings is a common finding in the South African labour market literature and partially reflects the extent of the income protection which unions are able to afford their members through their considerable collective bargaining power (Azam and Rospabé, 2007). The inclusion of this dummy variable also has the effect of again reducing the point estimates of the returns to educational attainment. Figure 6.6 illustrates this by showing the various earnings returns to education estimates from the models in Table A.4 graphically. While the estimated returns schedule becomes flatter as the regression model increases in complexity, it remains highly convex even for the specification in column (5).[16] This model is the most-fully specified of the estimations in Table A.4 and, as such, represents the baseline earnings returns specification to be used in the estimation of the sample selection corrected earnings returns in Section 6.4.

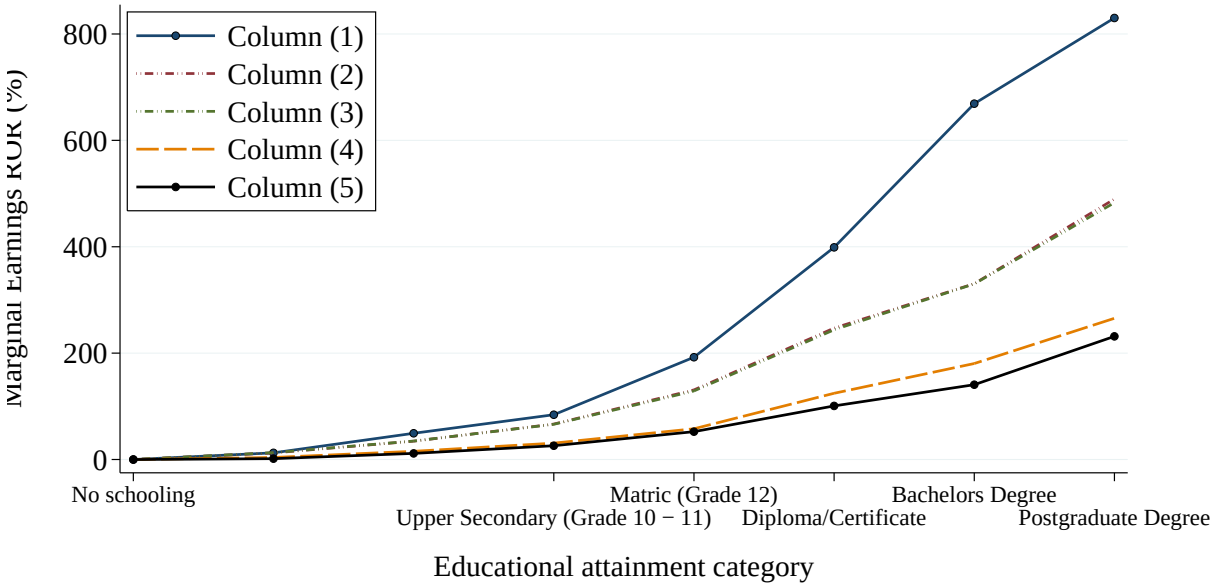
To test the hypothesis that the convexity in the earnings returns to education may be explained by school quality differentials, Table A.5 presents the results form the the earnings returns to education and school quality estimations. As is done in Section 6.2.1, the first column in the table displays the results from the baseline regression estimated for the full sample. Column (2) presents the baseline specification results when estimated using only those observations in the school quality sample. This reduces the estimation sample size from 4 452 to only 1 497 observations. There are other apparent differences in the coefficient estimates between the two estimations. However, an adjusted Wald test for cross-equation parameter equivalence reveals that these differences, with the exception of those between the coefficients on the Asian race dummy and the bachelors degree and postgraduate degree education dummies, are all statistically insignificant at the 10% level. There thus seems to be less heterogeneity between

[14] This is based on a English competency score of 6 for individuals who rated their ability to read and write English as "very well" vs a score of 0 for individuals who rated their English reading and writing abilities as "not very well at all". See Posel and Casale (2011) for a comprehensive analysis and discussion on the importance of language proficiency in the South African labour market.

[15] Calculated as $e^{\beta_{Union}} - 1$ from the log-linear estimation in column (5).

[16] Hypothesis tests reveal that $H_0 : ROR_{Diploma} = ROR_{Degree}$ for the estimation in column (5) cannot be rejected at below-10% levels of statistical significance. Despite the different point estimates, this implies that the rate of return to diploma and degree attainment levels are statistically the same. However, the returns to upper secondary, Matric, diploma/degree, and postgraduate attainment levels are statistically distinct at 5% significance or lower.

Figure 6.6: Estimated Uncorrected Earnings Returns to Educational Attainment



NOTES: The average marginal rares of return (ROR) calculated for each regression represent the *ceteris paribus* discrete average percentage change in earnings relative to the base category (no schooling): $\frac{\partial Pr(Earnings)}{(Educaton_j - Education_{base})} = (X\beta|Educ_j) - (X\beta|Educ_{base}) \ \forall j \neq base$. Each line corresponds to the estimates presented in one of the 5 columns of Table A.4.

the school quality sample and the baseline earnings returns estimation sample than between the school quality sample and the baseline employment returns estimation sample in Table A.2.

The school quality score is found to not be statistically significant when added to the baseline specification in model (3). The inclusion of the school quality measure also does not seem to impact significantly on any of the other coefficient estimates relative to the results for model (2).[17] As in Section 6.2.1, it is once again hypothesized that the school quality score variable could be redundant in the estimation because the racial dummy variables already sufficiently proxy for the inter-racial education quality differentials in South Africa. To test this hypothesis, the race dummies are excluded in model (4) of Table A.5 while the school quality measure is retained. The results show that the school quality measure is now highly statistically significant in the model, offering some support in favour of the aforementioned hypothesis.

To try and distinguish between the purely racial component of the unexplained differences in earnings between race groups and the component which may accrue to school quality differentials, Figure 6.7 displays the racial dummy coefficients from model (2) in Table A.5 alongside the average marginal school quality effect (AMSQE) for each race group from model (4). The AMSQE is a purely hypothetical construct and is calculated by multiplying the school quality coefficient estimate from regression (4) by the within-sample average school quality score for the race group under consideration. The AMSQE for Blacks is then subtracted from those of the

[17] Hypothesis tests show that none of the differences between the point estimates for models (2) and (3) are statistically significant at 10% significance.

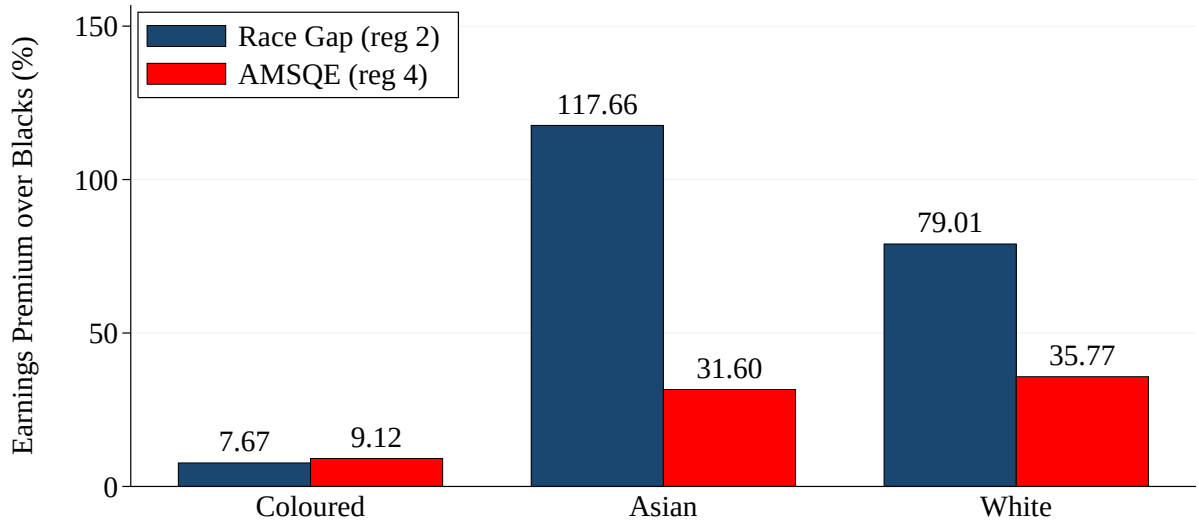
other race groups in order to make the estimates comparable with the racial dummy variables, where the Black race group serves as reference category. The graph suggests that it is plausible that a significant component of the White earnings premium over Blacks could be ascribed to the fact that Whites in the estimation sample have received far better quality schooling, on average, than Blacks. Similar results follow for Coloureds and Indians, albeit to differing degrees.

The estimates in Figure 6.7 would thus suggest that school quality differentials may explain a significant portion of the unexplained observed earnings differences between different race groups in South Africa. Put differently, it may be the case that a substantial part of South Africa's racial earnings inequality does not accrue from labour market discrimination, as is often believed, but rather from racial inequalities in the education system. Moreover, the point estimate for the school quality variable in regression (4), on which the AMSQEs in Figure 6.7 are based, is likely to be biased due to both sample selectivity and measurement error. If it is the case that individuals who have benefited from better quality schooling are also more likely to be employed and, consequently, to be included in the non-zero earnings sample, it is to be expected that the extent of the inter-racial variation in school quality would be less in the earnings sample than in the population of working age. Coupled with the fact that the nature of the school quality measure used here tends to exclude individuals towards the lower end of the educational attainment distribution - where quality differentials are not necessarily any less severe - from the estimation sample, it is likely that the the estimates in Figure 6.7 actually underestimate the extent to which average racial school quality differentials may explain labour market earnings gaps. However, in the absence of a more representative estimation sample and a better measure of school quality, this remains only conjecture. The measurement issue cannot be addressed directly in this study, but an attempt is made to account for the sample selection effects on these estimates in Section 6.4.

The results for the baseline uncorrected earnings returns to educational attainment and numeracy estimations are presented in Table A.6. The presentation of the results follow broadly the same structure as Table A.5. There is again a marked reduction in the estimation sample size when estimating the baseline earnings returns regression for the numeracy sample. Only 885 of the original 4 452 observations remain when those individuals who do not have numeracy scores available are excluded from the estimation. This censoring of the estimation sample has pronounced effects on most of the regression estimates. The returns to education for the numeracy sample appear to be significantly greater than for the full earnings sample. On closer inspection, however, it is revealed that it is only the differences in coefficients for the diploma/certificate and degree educational attainment levels that are statistically significant at 10% between models (1) and (2). This reveals that, similar to the case where the baseline employment returns regression was estimated for the numeracy sample, the censoring of the number of observations in the numeracy sample leads to large standard errors and poor estimate precision.

The inclusion of the numeracy score variable in column (3) does not change the educational

Figure 6.7: Uncorrected Unexplained Racial Earnings Gaps and Average Marginal School Quality Effects (AMSQE) for the School Quality Sample (Table A.5)



NOTES: Race gaps represent the racial dummy coefficient estimates for the school quality sample from regression 2 in Table A.5. AMSQEs are calculated by multiplying the school quality coefficient estimate from regression 4 in Table A.5 by the within-estimation-sample mean school quality score for the race group under consideration.The AMSQE for Blacks is subsequently subtracted from the AMSQE for each of the other race groups. The Black race group thus represents the reference category in the graph.

attainment coefficient estimates significantly, nor any of the estimates for the other control variables. The numeracy score is itself only statistically significant when the educational attainment dummies are excluded from the model and then only at the 10% level. To some extent, these findings echo those from the previous section where it was argued that the small estimation sample size for the numeracy sample may insurmountably constrain the scope for obtaining reliable parameter estimates for the earnings returns to educational attainment in conjunction with the earnings returns to numeracy levels. One important difference here is that the educational attainment dummies remain statistically significant in the earnings returns estimation for the numeracy sample. However, in the absence of any significant changes in these coefficient estimates as a direct result of the inclusion of the numeracy score measure, one must either conclude that numeracy, in the way that it is measured here and the sample within which it is captured, does not account for any of the ability bias in returns to educational attainment estimates or that there simply is no significant ability bias present in these estimates.

The uncorrected analysis above offers some mixed results. It is clear that the censoring of estimation sample due to the small number of observations with numeracy or school quality data becomes problematic for the estimation of the parameters of interest. This is particularly true given that the sample selection processes for the numeracy and school quality measures appear to homogenise the estimation sample in terms of educational attainment levels and other human capital measures. However, some tentative conclusions can be drawn from the findings. Fist,

the preliminary results seem to suggest that neither school quality nor numeracy has a significant role to play in explaining the convex nature of the employment and earnings returns to educational attainment in South Africa. Second, some support is offered in favour of the hypothesis that a non-trivial component of unexplained interracial earnings gaps may be explained by inter-racial school quality differentials, although the evidence is somewhat circumstantial owing to the imperfect measurement of the school quality concept and the selectivity of the school quality sample used here. Lastly, it is difficult to gauge the relative importance of educational attainment, numeracy and school quality for labour market outcomes from the estimates presented thus far since the potentially contaminating effects of sample selectivity on those estimates cannot be ruled out yet. In order to provide a more definitive answer on this matter and assess the validity of the aforementioned preliminary conclusions, it is therefore necessary to explicitly control for potential endogeneity as a consequence of non-random selection.

6.3 Selection Estimation

The multivariate analysis now proceeds to the intermediate stage in the estimation strategy which entails the modelling of the four selection processes identified in Section 6.1 that determine whether observations are included in the various labour market outcome estimation samples. Specifically, the selection processes to be modelled are: labour force participation as selection into the participant sample, employment as selection into the earnings sample, being matched to a school as selection into the school quality sample, and numeracy test module participation as selection into the numeracy sample. The objective of these estimations is to establish the first-step selection equations which need to be used in order to estimate the sample selection corrected employment and earnings returns in the final section of this chapter.

As explained in Section 4.2.2, obtaining unbiased estimates when using the Heckman (1979) MLE procedure requires, among other things, careful modelling of the first-stage selection process. The goal in this stage of the analysis is therefore to model the respective selection processes as accurately as possible, while taking care not to undermine the primary objectives for the multivariate analysis. Given the data limitations, this goal most likely necessitates a compromise between the adequacy of the first-step selection equations and the adequacy of the second-step outcome equations. To clarify, each first-step selection equation should be modelled in such a manner that the large number of missing values on certain variables does not unduly limit the number of observations available for the second-step outcome estimation, or require important second-step covariates to be excluded from the outcome estimation in order to avoid high multicollinearity and improve parameter identification. This latter requirement simply means that the explanatory variables for the baseline outcome equations identified in the previous section should not have to be excluded from outcome estimations just to ensure a suf-

ficient number of exclusion restrictions in the selection equations. Together, these requirements place certain practical restrictions on the variables that can be used to model selection.

Given that the interest of this chapter lies in obtaining estimates of the labour market returns to educational attainment, school quality, and numeracy, the sole purpose of modelling the selection processes is to assist in reducing the bias of the outcome estimates. In other words, the results of the selection estimations are interesting only insofar as they provide an indication of how adequately a particular selection process is being modelled. In general, the estimation results that follow are therefore discussed only in superficial terms. The baseline employment returns model in column (4) of Table A.1 is a sufficiently adequate representation of the selection process that determines whether individuals have non-zero earnings observable. Consequently, this model will be used as the baseline selection equation for the labour market earnings outcome and, given that it has already been discussed in the previous section, will not be discussed again below.

Table A.7 presents the baseline models for the other three selection processes. Many of the household structure and physical well-being variables that are included in the baseline employment returns regression actually influence the probability of being employed indirectly through their associations with the likelihood of labour force participation (LFP). When modelling selection into the participant sample, it is therefore appropriate to use these variables as covariates in the LFP equation instead of the employment outcome equation. The LFP specification thus includes the household head status, marital status, health status and disability dummy variables previously included in the employment returns estimations.

The coefficient estimates indicate that the propensity to participate in the labour market initially increases with age. After the age of about 37, however, individuals increasingly become less likely to be engaged in economic activity. Similar to the original result for the uncorrected employment returns estimations, LFP is found to be increasing in educational attainment and physical well-being and is negatively affected by having a disability or being enrolled in education. Despite the significant rise in female labour force participation rates over the past 15 years, females - married females in particular - are still significantly less likely to participate than their male counterparts. There also appears to be a non-linear relationship between LFP and household size according to which LFP is predicted to fall as households become larger, but then begins to rise again for households that exceed four members. Lastly, a dummy variable which indicates whether an individual resides in a household that receives some form of social grant income is included in the regression to allow for the possibility that income received through social grants could discourage LFP. While the negative and statistically significant coefficient on this dummy variable appears to support such a hypothesis, the marginal effects estimates reveal that individuals from households that receive social grants may on average be only 6 *pp*

less likely to participate in the labour market than those who do not.[18]

From the discussion of the NIDS school quality variable in Section 3.3 it should be clear that an individual could only be matched to school quality data (and thus be included in the school quality sample) if three conditions were met: first, the individual must have been willing to provide information on the educational institution which he/she attended; second the individual needed to be able to provide sufficiently accurate and detailed information in order for the institution that was identified to be matched with a school in the South African schools registry; and third, matric performance data must have been available for the school in question. The probability of being included in the school quality sample should thus equal the joint marginal probability of the three conditions being met simultaneously. If it is assumed, for the sake of simplicity, that the process determining whether or not Matric performance data is available for a school is completely random, then modelling selection into the school quality sample reduces to modelling the probability that individuals are willing and able to provide accurate information on the educational institution(s) which they attended.[19] It is this probability that is modelled in the school quality match selection equation in Table A.7.

The results from the estimation confirm that better-educated and younger individuals are more likely to be included in the school quality sample. Coloured individuals also have a significantly higher probability of being matched to school quality data than Black individuals. The home language dummy variables further control for ethnic differences and show that individuals belonging to some of South Africa's indigenous language groups have a higher likelihood of being included in the school quality sample than English-speaking individuals.

The *father educ info* and *mother educ info* dummy variables in the regression indicate whether individuals provided information on their parents' levels of educational attainment. The outcomes on these measures partially reflect an individual's willingness and ability to divulge detailed education-related information to fieldworkers. The coefficient estimates on the two variables are both statistically significant and positive, suggesting that individuals who are able to provide information of this nature may be more likely to also provide information which would ensure that they can be linked to a school. The significant and positive coefficient on the *currently enrolled* variable provides some support for the hypothesis that individuals who are currently enrolled in education may have an advantage in recalling details about the high

[18] The findings from the literature investigating the impact of social grants on labour market search activities in South Africa is inconclusive. For example, while Posel *et al.* (2004) find that the South African social pension facilitates search activities for household members, Betrand *et al.* (2003) find that the social pension actually reduces labour force participation rates.

[19] Even if it was the case, for example, that Matric performance data is less likely to be available for those schools identified by older individuals - schools that may no longer exist or have had their names changed - it is unlikely that there would have been a large number of cases of missing Matric performance data. It is therefore safe to assume that the probability that the third condition is met, conditional on the satisfaction of the first two conditions, is close to unity and has a negligible impact the number of individuals ultimately excluded from the school quality sample.

school that they attended. Lastly, a *household school quality match rate* variable measuring the proportion of other household members that could successfully be linked to school quality data is included in the estimation.[20] The inclusion of this variable allows one to gauge the extent to which the structure of, prevailing culture in, and behaviours of the members of a household influence individuals' decisions to respond to certain questions. In the present context, it is conjectured that individuals may be more capable of providing school quality information and also be more inclined to do so if other members in the household are able to offer them assistance or are observed to divulge school information themselves. The estimation results suggest that this is indeed the case. The probability of being included in the school quality sample increases significantly as the proportion of other household members that are successfully linked to school quality data rises.[21]

The probability of participating in the numeracy test module is modelled in the right-most section of Table A.7. The model specification used is based on the descriptive analysis in Section 3.2.2 and the estimations conducted in Du Rand *et al.* (2010) [22] Given the considerable response burden associated with participating in a voluntary cognitive assessment test, it is not surprising that the estimation results show the probability of test participation to be increasing in educational attainment, health, and emotional well-being. The curriculum-based content of the numeracy test also means that younger respondents who are either still in formal education or only recently left education should find the prospect of participating in the test less daunting than older individuals, all else being equal. This supposition is supported by the results which show that the probability of participating in the test decreases as individuals get older.[23]

The numeracy test module occurred near the end of the NIDS survey questionnaire. Assuming that marginal response effort is increasing in the amount of time spent on answering a survey, it should be expected that respondents would become increasingly unwilling to engage in further voluntary survey participation, the longer the amount of time they have already spent on the survey. The *time before test* variable reflects the time that elapsed before the NIDS numeracy test was enumerated to respondents. The estimated coefficients on this variable reflects the extent to which respondent fatigue may deter individuals from participating in cognitively challenging assessment tests that form part of general household survey questionnaires.

Took measurements is a dummy variable indicating whether an individual submitted to having

[20] Here "other" refers to all household members excluding the individual under consideration.

[21] A quadratic term for the household school quality match rate is included to reflect the fact that there are diminishing marginal returns to increases in the proportion of matched household members.

[22] Given that the numeracy test participation decision is already discussed comprehensively in Du Rand *et al.* (2010), and the fact that understanding the underlying factors that determine whether individuals participate in numeracy tests is not a primary objective in this study, the numeracy sample selection model in this section is not discussed extensively.

[23] Since it is only the population of working age that is considered in the analysis, the estimated turning point for the age variable, after which the probability of participating in the test would presumably begin to increase again, falls outside the feasible age range. This implies that the probability of being selected into the numeracy sample is, in practice, strictly decreasing in age.

biometric measurements taken in the survey section immediately preceding the enumeration of the numeracy test module. Whether or not individuals were willing to do so offers some indication of both their willingness to subject themselves to assessment and participate in voluntary components of the survey and the degree of their perseverance in answering the survey. On this basis, one would expect individuals who had their measurements taken to also have been more willing to participate in the numeracy test module. The estimation results suggest that this is indeed the case, showing a statistically significant and positive coefficient on the *took measurements* dummy. The final variable in the model, the *household test response rate*, is similar to the *household school quality match rate* variable included in the school quality selection estimation. The variable now measures the proportion of other household members who were eligible to participate in the numeracy test module and opted to do so. It is again shown that individuals have a higher probability of participating in the numeracy test if other household members also choose to do so.

The model fit statistics for the three selection estimations in Table A.7 and the employment selection estimation in Table A.1 indicate that the models respectively predict the observed outcomes for LFP, employment, school matching, and numeracy test participation with reasonable accuracy. When compared to the baseline outcome estimation specifications in Section 6.2, each selection equation is also seen to include a number of valid exclusion restrictions. All four selection estimations thus appear to satisfy the theoretical requirements for use in the sample selection correction estimation procedure described in Section 4.2.2 and hence collectively constitute the set of first-stage selection equations on which the estimation of the corrected labour market returns to educational attainment, school quality and numeracy in the next section are based.

6.4 Sample Selection Corrected Estimation

Tables A.8 - A.11 present the results from the various sample selection corrected estimations of the employment and earnings returns to educational attainment, school quality and numeracy. The selection equations used in these estimations correspond to those presented in the previous section while the outcome equations correspond to those in Section 6.2.[24] As explained above, the specification of the baseline employment returns outcome model is altered to allow for variables that influence the probability of being employed indirectly via the LFP channel to be included in the LFP selection rather than the employment outcome regression. The standard baseline employment returns specification in column (4) of Table A.1 is therefore only used to model selection into the earnings sample in the corrected earnings returns estimations. The

[24] Given that the results for the respective selection equations are already presented in full in Tables A.7 and A.4, they are omitted from the presentation of the corrected employment end earnings returns estimations in Tables A.8 - A.11.

output tables are structured to align the presentation of the estimation results with the structure followed in the third-stage of the estimation strategy outlined in Section 6.1.[25]

6.4.1 Corrected Employment Returns

Column (i) in Table A.8 shows the estimation results for the uncorrected adjusted employment returns model when estimated for the unrestricted working-age sample. An adjusted Wald test confirms that the coefficients on the educational attainment dummies in this model are statistically different from those estimated for the baseline employment returns model in column (4) of Table A.1. Equation (ii) re-estimates the model for labour force participants only, producing estimates of which the vast majority are statistically different from those in column (i) at at least 5% significance. These changes show that the estimation results are highly sensitive to the model specification and the estimation sample used.

Column (1) presents the results for the employment returns to educational attainment estimation when correcting for selection into LFP. Hypothesis tests show that none of the estimated coefficients for this model are statistically different at conventional levels of significance from those for the equivalent uncorrected estimation in column (ii).[26] The reason for the lack of any statistically significant differences between the estimates is revealed in the statistical insignificance of the cross-equation error correlation terms. The estimate for $\rho_{Participation}$ in column (1), measuring the correlation between the error terms from the first-stage LFP selection equation and the second-stage employment outcome equation, shows that the extent of the correlation between the first and second-stage error terms is not statistically different from zero.[27] This is an important result since it suggests that any commonality between the unobservables that influence selection into the participant sample (i.e. LFP) and the unobservables that influence the employment outcome itself is negligible once the explanatory factors that have been included in the estimation are accounted for. Conditional on the validity of the model specification and the assumptions underlying the estimation procedure implemented, it follows that one may reject the hypothesis that the coefficient estimates in the employment outcome equation are biased due to endogenous selection into the participant sample. In other words, the estimates obtained using the sample selection correction procedure are not any less biased than what would have been the case if the selection and outcome equations were estimated separately.

[25] All regressions are estimated in Stata/SE 11.2 using either the *svy: heckprob* command for the estimation of the corrected employment returns or the *svy: heckman* command for estimation of the corrected earnings returns. (StataCorp, 2009*b*).

[26] In the remainder of this paper, *conventional levels of significance* and *reasonable levels of significance* are used interchangeably to refer to the 10%, 5% and/or 1% levels of statistical significance.

[27] In order to constrain ρ to the [-1,1] interval and maintain numerical stability during the ML optimization, Stata actually estimates the inverse hyperbolic tangent of ρ, atanh ρ, and not ρ itself. The estimate of atanh ρ can then be transformed back into an estimate of ρ. However, since $atanh\,(0) = 0$, a test for H_0 : atanh $\rho = 0$ is equivalent to a test for $H_0 : \rho = 0$. (StataCorp, 2009*a*, pp. 650 - 651)

An inspection of the rest of the estimations in Table A.8 reveals that none of the cross-equation error correlations are statistically significant. There is thus no statistical evidence in these models to suggest that failure to control for either selection into LFP or selection into the school quality sample biases the employment returns to educational attainment and school quality estimates. This result appears to be encouraging in the sense that it suggests one may now have greater confidence in the fidelity of the estimates obtained in the uncorrected employment returns models in Table A.2. However, it is worth noting that the absence of any statistical proof of selection bias in Table A.8 does not necessarily prove that the selection processes involved are completely exogenous. Given the selection correction procedure's sensitivity to the model specification and any violations of the underlying distributional assumptions, one cannot rule out the possibility that the results which obtain do so due to model misspecification or the small number of uncensored observations in the outcome estimation samples. The alternative would be to conclude that the underlying characteristics of those individuals included in the employment and school quality samples do not differ from those of their excluded counterparts in unobserved ways when considered in terms of the factors that influence the probability of being employed.

The models in columns (2) to (7) of Table A.8 restrict the outcome estimation sample to those observations that fall within the school quality sample. When compared against the results in column (1), the coefficient estimates are shown to change in a manner similar to that observed for the uncorrected estimates in Table A.2. Hypothesis testing shows that none of the employment returns to educational attainment estimates (or any of the other coefficients) change in a statistically significant way in response to the inclusion or exclusion of the school quality variable or the implementation of the selection corrections. What differences there are between the estimates for the various corrected models actually are due to the differences in the number of uncensored observations available for the second-stage outcome estimations.

The estimates of the returns to Matric, diploma, and degree attainment levels become insignificant when the outcome estimation sample is constrained to include only overlapping school quality and participant sample observations in columns (2) to (4), but are significant again when the LFP constraint is lifted in columns (5) to (7). The impact of the school quality score variable is only statistically significant when the race dummies are excluded in specification (4), and then only at the 10% level. However, the most likely reason for the significance of this variable here and its insignificance in regression (4) of Table A.2, is the change in model specification and estimation sample.

A comparison of the differences in the coefficient estimates among the various corrected models in Table A.8 and the differences between these estimates and the uncorrected estimates in Table A.2 highlights an important issue pertaining to the modelling of labour market returns in South Africa. The results suggest that the estimation outcomes may be more sensitive to the inclusion

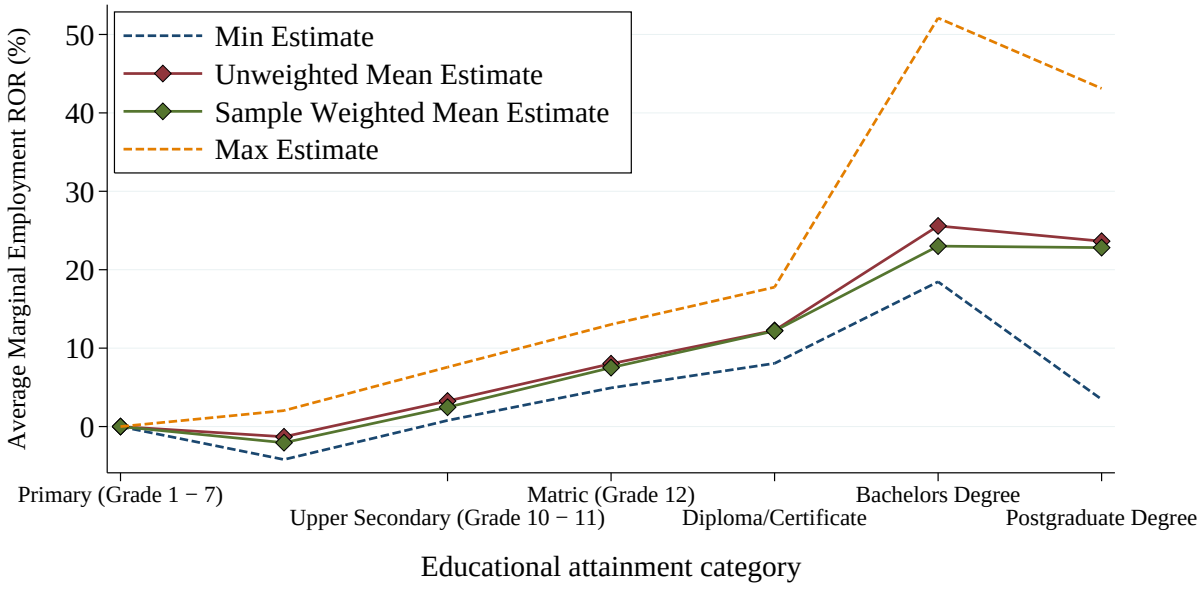
or exclusion of certain explanatory variables or the censoring of the estimation sample than to the implementation of different estimation procedures and corrections. This illustrates that the gains brought about by the use of sample selection correction procedures may be limited if models are not correctly specified.

Table A.9 displays the results from the corrected employment returns to educational attainment and numeracy estimations. The estimates in column (1) are the same as those in column (1) of Table A.8 and correspond to the LFP-corrected adjusted employment returns model. None of the estimates of $\rho_{Participation}$ are statistically different from zero at any reasonable level of significance. This finding adds further support for the rejection of the hypothesis that failing to explicitly control for selection into LFP may bias employment returns estimates. In contrast to what is observed for the uncorrected returns to numeracy estimations in Table A.9, the magnitudes of the returns to education actually increase substantially when the outcome estimation sample is limited to include only observations in the numeracy sample, as in columns (2) and (3).[28] However, the coefficient on the numeracy score variable only becomes significant when the educational attainment dummies are excluded as in model (4). Moreover, the marginal effects estimates for this model show that the impacts of numeracy may be negligible, with a one standard deviation increase in numeracy being associated with only a 3.15% rise, on average, in the probability of being employed, *ceteris paribus*.

The estimates for the cross-equation error correlations between the first-stage numeracy sample selection equation and the second-stage employment outcome equation in models (5) and (6) are statistically significant at the 10% level. This is the first result to provide some support for the notion that it may be necessary to explicitly correct for sample selection in the estimations of the South African labour market returns to numeracy in this study. Specifically, a statistically significant $\rho_{Numeracy}$ suggests that there are common unobservables in the first- and second-stage estimation error terms which not only influence the probability of being selected into the numeracy sample but also the probability of being employed and which have not yet been controlled for in the model specification. Given that the estimation sample for model (6) corresponds roughly to the estimation sample for model (3) in Table A.3, the model results thus offer an opportunity to gauge the extent to which the sample selection correction procedure may have influenced the returns to educational attainment and numeracy estimates. The adjusted Wald tests do indeed suggest that one may reject the hypothesis that the corrected and uncorrected estimates of the returns to Matric, diploma, and degree attainment levels are statistically equivalent at at least 10% significance. However, a simple robustness test reveals that the reason for the statistically significant differences in the aforementioned corrected and uncorrected estimates is not the sample selection correction procedure, but rather the differences in the

[28] The inflated returns estimates for the bachelors degree dummies in equations (2) and (3) in Table A.9 are caused by the fact that the 25 labour force participants with bachelors degrees in the estimation sample are all employed. In other words, LFP perfectly predicts the employment outcome for these observations, thus hindering the accurate estimation of the employment returns to having a bachelors degree.

Figure 6.8: Average Marginal Employment Returns to Educational Attainment: Summaries



NOTES: The max, unweighted mean, sample weighted mean and min estimates presented in the graph are composite summary measures that correspond to the maximum, mean, estimation-sample-weighted mean, and minimum average marginal employment returns to educational attainment estimates calculated for the following 9 regressions: (4) in Table A.1; (3) in Table A.2; (3) in Table A.3; (i), (3) and (6) in Table A.8; and (3) and (6) from Table A.9. The average marginal effects calculated for each regression represent the discrete average percentage change in the probability of being employed relative to the base category (completed primary education), evaluated over the observed values of the other regression covariates in the estimation sample: $\frac{\partial Pr(Employment)}{(Educaton_j - Education_{base})} = (\Phi(X\beta)|Educ_j) - (\Phi(X\beta)|Educ_{base}) \ \forall j \neq base$

specifications used in model (3) of Table A.3 and model (6) of Table A.9.[29] By implication, it is again unclear whether the sample selection correction procedures applied in the estimation of the models in Table A.9 actually yield any significant gains in terms of reducing the bias in the estimates of the employment returns to the human capital measures in the NIDS data relative to the uncorrected estimations in Section 6.2.1.

The discussion of the estimation results above alludes to the fact that the estimates of the employment returns to educational attainment in South Africa may vary significantly depending on the sample used in the estimation and the way in which the estimation model is specified.[30] Focussing on one specific set of estimates, as many studies do, would therefore not be prudent. Instead, it is preferable to establish hypothetical maximum and minimum confines within which the estimates are reasonably likely to fall.[31] Figure 6.8 plots the overall minimum, mean, and maximum estimates of the average marginal employment returns to educational attainment cal-

[29] This robustness test entails re-estimating model (3) from Table A.6 using the adjusted employment outcome model in column (i) of Table A.8 with the numeracy measure included. The adjusted Wald tests are then repeated to test for statistically significant differences between the estimates in the new model and those for model (6) in Table A.9.

[30] The estimates are, of course, also likely to differ depending on the data that is used in the analysis.

[31] These confines would then represent a type of pseudo confidence interval, though not one which has any defined statistical meaning.

culated for some of the corrected and uncorrected regressions presented in this Chapter.[32] The differences in the respective estimates capture the effects of changes in the model specification and estimation sample, and the impacts of the inclusion/exclusion of the numeracy and school quality measures and the estimation approaches used.

While the difference between the minimum and maximum returns estimates in the graph remains relatively constant up to diploma/certificate levels of attainment, the estimates diverge substantially for post-diploma attainment levels. One reason for this divergence is that the proportion of employed individuals at each level of educational attainment differs substantially between the respective estimation samples used in the regressions.[33] The unweighted and estimation-sample-weighted arithmetic means, taken over the range of estimates from the various regressions, are also plotted on the graph in an attempt to gauge the general structure of the employment returns to educational attainment profile in South Africa from the estimations in this study. The calculation of the estimation-sample-weighted mean is based on the rationale that the results obtained from small estimation samples are theoretically more likely to be biased or estimated with less precision than those from large estimation samples. This measure thus assigns a smaller weight to the estimates obtained from the school quality and numeracy samples than to the estimates obtained for the full working-age sample.

A number of conclusions can be drawn from the plots in Figure 6.8 and the discussion of the employment returns estimation results above. First, although not extensively discussed, the estimation results suggest that the employment returns to race in South Africa are statistically negligible once inter-racial differentials in educational attainment levels have been taken into consideration. In other words, there does not appear to be any unexplained advantage for one race group over another in terms of the probability of being employed.[34] Second, there is no robust evidence to suggest that school quality, in the way that it is measured here, has a statistically significant impact on the probability of being employed once other standard observables have been controlled for. Similarly, while some of the results suggest that numeracy skills have a statistically significant and positive effect on the employment outcome, the marginal effects estimates show that the magnitude of this effect is, at best, trivial in relation to the effects of other factors like educational attainment.

[32] The results from the following regressions were used to calculate the minimum, mean, and maximum summary measures: (4) in Table A.1; (3) in Table A.2; (3) in Table A.3; (i), (3) and (6) in Table A.8; and (3) and (6) from Table A.9. Collectively, these regressions are representative of the different samples, model specifications and approaches used in the etimations of employment outcomes in South Africa. The respective estimates are assigned equal weight in the calculation of the arithmetic mean.

[33] The maximum estimate for individuals with bachelors degrees corresponds to the estimate for regression (3) in Table A.9 which, as noted above, is inflated because all of the individuals with bachelors degree attainment levels in the estimation sample are employed.

[34] Where the coefficients on the race dummies are statistically significant, this is because the racial composition of the estimation sample deviates strongly from the racial composition of the full population of working-age sample. See, for example, the results for models (2), (3), (5) and (6) in Table A.8 and models (2) to (7) in Table A.9.

The majority of the estimations report statistically significant employment returns to educational attainment. This finding also appears to be reasonably robust to the non-random censoring of certain estimation samples.[35] However, these significant returns are shown to manifest almost exclusively at higher, post-Matric levels of attainment. As such, the convex structure of the employment returns to attainment schedule appears to be invariant to the inclusion of the numeracy and school quality measures. Moreover, the actual magnitudes of the returns to educational attainment are found to be largely unaffected by the inclusion of these measures.

Lastly, there is circumstantial statistical evidence that some unobserved factors that influence selection into the numeracy sample may also influence employment outcomes. However, the *heckprob* procedure, implemented to explicitly control for any bias in the employment returns estimates that result firom endogenous selection into the numeracy sample produces estimates which are not statistically different from those obtained under OLS. In fact, neither the corrections for selection into LFP or selection into the numeracy or school quality samples are found to produce substantively different results from those for the corresponding uncorrected estimations.

6.4.2 Corrected Earnings Returns

Table A.10 presents the results for the sample selection corrected earnings returns to educational attainment and school quality regressions. Column (1) shows the estimates for the baseline earnings returns estimation in column (5) of Table A.4, when correcting for selection into employment using the baseline employment returns to education estimation in column (4) of Table A.1. Similar to the findings in Section 6.4.1, the adjusted Wald tests reveal that one cannot reject, at any reasonable level of significance, the hypothesis that the coefficients in this corrected model are equal to those for the uncorrected model (when estimated for the same sub-sample). Again, this result is partly explained by the fact that the correlation coefficient between the first and second-stage equation error terms, $\rho_{Employment}$, is found not to be statistically different from zero. However, assuming that model (1) is correctly specified, this would imply, somewhat counter-intuitively, that employed and unemployed individuals in South Africa are not characteristically different from one another in terms of the unobservables that influence earnings outcomes. More precisely, it suggests that the covariates included in the earnings outcome estimation account for all of the earnings capacity relevant dimensions along which the employed differ from the unemployed.[36]

[35] There are, of course, cases where the returns to educational attainment appear to not be statistically significant (see models (2) and (3) in Table A.3 and models (2) to (4) in Table A.3.

[36] Several different model specifications for each of the estimations in Table A.10 were tested to examine the robustness of the finding that $H_0 : \rho_{Employment} = 0$ cannot be rejected at any reasonable level of statistical significance. In general, this finding proves to be robust to the inclusion of additional explanatory variables in the second-stage earnings outcome specification. However, the inclusion of additional exclusion restrictions in

Column (2) shows the results for model (1) when the outcome equation is estimated only for those individuals included in the school quality sample. The vast majority of the coefficient estimates for this model appear to differ significantly from those obtained in the equivalent uncorrected estimation in column (3) of Table A.5. The magnitudes and statistical significance of the employment-corrected earnings returns to educational attainment estimates, for example, appear to be substantively different from what is found in the uncorrected estimation.[37] However, cross-equation parameter equivalence tests reveal that one cannot, at any conventionally acceptable level of statistical significance, reject the hypothesis that the two sets of coefficient estimates are statistically equivalent. This seemingly peculiar result obtains due to the inflated confidence intervals surrounding the point estimates for the corrected model. In essence, the implementation of the selection correction procedure imposes additional restrictions on the second-stage earnings outcome equation which undermines the precision with which the coefficients are estimated. This lack of precision is also observable in the statistically insignificant estimate for $\rho_{Employment}$ in model (2). Considered in isolation, the magnitude of the estimate would suggest that the correlation between the residuals for first-stage employment selection equation and the residuals for the second-stage earnings outcome equation is negative and large. However, because the confidence intervals for this estimate are so wide, the hypothesis that it is not statistically different from zero cannot be rejected.

The included school quality measure in the employment-corrected earnings return estimation in column (3) is shown not to be statistically significant and not to have a statistically significant impact on the other estimated coefficients. In fact, the estimated coefficients and statistics for the employment-corrected earnings returns model only change when the race dummies are excluded in model (4). Similar to the results for the uncorrected estimations, the school quality variable is shown to be statistically significant at the 1% level when the model is estimated without controls for race, suggesting once again that race may be a strong proxy for school quality in South Africa. The magnitude of the school quality effect is, however, smaller than in the uncorrected estimation in column (4) of Table A.2. Given that these two estimations are conducted on virtually the same estimation samples, this shows that the correction for selection into employment attenuates the partial effect of school quality on earnings.[38]

When compared to the results for models (1) to (3) in Table A.10, it seems strange that the estimate of $\rho_{Employment}$ in model (4) is suddenly found to be statistically significant at the 1% level. However, this result should not necessarily be interpreted as proof that selection into

the first-stage employment outcome estimation produces mixed results, leading to the rejection of the aforementioned hypothesis in some cases and failure to reject in others. With no clearly discernible pattern emerging, it was decided to keep the corrected model specification consistent with the specifications used in the respective uncorrected earnings and employment returns models in Section 6.2.1.

[37] The phrase "employment corrected" is used here as shorthand for estimations where some form of correction has been made in an attempt to control for selection into employment.

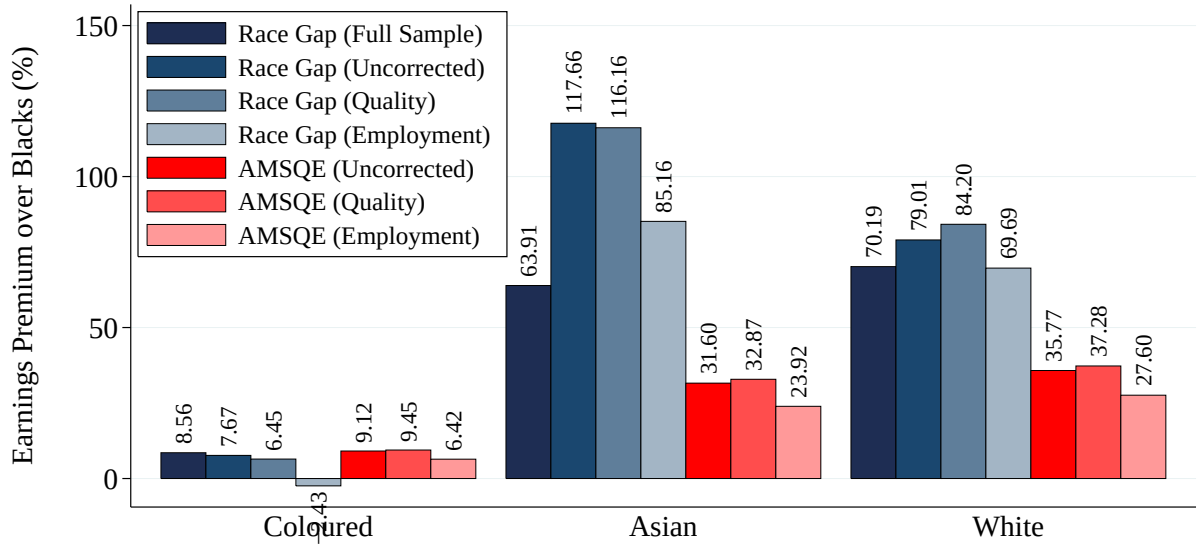
[38] The difference between the two school quality coefficients in the employment-corrected model and the uncorrected model is statistically significant at the 10% level.

employment is endogenous to the earnings outcome in equation (4) while being exogenous in models (1) to (3). Instead, the abrupt change in the significance of ρ illustrates how sensitive the Heckman (1979) MLE approach can be to small changes in the model specification or violations of the distributional assumptions on which it is based, and how this sensitivity can lead to unstable estimation results. Specifically, the omission of the racial dummy variables from the outcome equation in model (4) effectively "converts" the race dummies in the selection equation into three additional exclusion restrictions. The observed statistical significance of the cross equation error correlation coefficient could therefore be attributed to the fact that these additional exclusion restrictions allow for better identification and more precise estimates of the model statistics.

Models (5) to (7) in Table A.10 show the results for the same earnings outcome equations as in columns (2) to (4), but now controlling for selection into the school quality sample. None of the estimates for $\rho_{Quality}$ are found to be statistically different from zero, suggesting that it is unlikely that the uncorrected estimates of the earnings returns to education in models (3) and (4) in Table A.5 are biased as a result of endogenous selection into the school quality sample. Hypothesis testing also shows that the returns to educational attainment estimates in models (5) and (6) in Table A.8 and model (3) in Table A.5 are statistically equivalent. In other words, neither the inclusion of the school quality measure nor the correction for selection into the school quality sample alters the coefficient estimates for this particular estimation sample. The coefficient on the school quality variable is again only statistically significant in model (7) when the race dummies are omitted from the estimation. All of the coefficient estimates for this model are also found to be statistically equivalent to those in model (4) of Table A.5.

While there appears to be some minor differences in the earnings returns to race estimates for the different models in Table A.5, the race effects within each estimation sample are shown to be statistically unaffected by the inclusion of the school quality measure. To provide some indication of the potential extent to which school quality differentials may contribute to racial earnings differentials in South Africa, the earnings returns to school quality estimates from models (4) and (7) in Table A.5 are again used to construct hypothetical AMSQEs for each race group. Figure 6.9 illustrates these estimates graphically alongside the uncorrected AMSQE estimates calculated in Section 6.2.2 and the various uncorrected and corrected unexplained returns to race estimates. The relative magnitudes of the sample selection corrected estimated AMSQEs and unexplained racial earnings gaps in the figure appear to be broadly the same as those found for the uncorrected AMSQEs and unexplained racial earnings gaps. The corrected estimates thus neither detract from nor add substantively to the conclusions that could be drawn from Figure 6.7 in Section 6.2.2. In other words, despite the failure to find significant earnings returns to school quality in conjunction with returns to race, there is some circumstantial evidence that a non-trivial component of the unexplained average racial earnings gaps in South Africa may be attributed to inter-racial school quality differences. Moreover, given that the representativeness

Figure 6.9: Unexplained Racial Earnings Gaps and Average Marginal School Quality Effects (AMSQE)



NOTES: Race gaps represent the respective racial dummy coefficient estimates from model (5) in Table A.4 (Full sample), (2) in Table A.5 (Uncorrected), (2) in Table A.10 (Employment), and (5) in Table A.10 (Quality) . AMSQEs are calculated by multiplying the school quality coefficient estimate from regression (4) in Table A.5 (Uncorrected), (4) in Table A.10 (Employment), and (7) in Table A.10 (Quality) by the within-estimation-sample mean school quality score for the race group under consideration.The AMSQE for Blacks is subsequently subtracted from the AMSQE for each of the other race groups. The Black race group thus represents the reference category in the graph. With the exception of the race gaps for the full sample, all estimates are for the school quality sample.

of the school quality measure used here appears to be biased in favour of individuals towards the upper end of the educational attainment distribution, it remains likely that the picture portrayed in Figure 6.9 still underplays the role of school quality differentials in explaining observed racial differentials in certain labour market outcomes.

Table A.11 presents the results for the various sample selection corrected earnings returns to educational attainment and numeracy estimations. Column (1) reports the results from the same employment corrected earnings returns estimation as shown in column (1) of Table A.10. As already discussed above, the results for this model are statistically equivalent to those for the uncorrected model in column (1) of Tables A.5 and A.11. Models (2) to (7) restrict the earnings outcome estimation sample to include only those observations in the numeracy sample with (2) to (4) correcting for selection into employment and (5) to (7) correcting for selection into the numeracy sample. The estimates for $\rho_{Employment}$ in models (2) and (3) are both shown to not be statistically different from zero. It is therefore not surprising that the pairwise adjusted Wald tests show the one cannot reject, at any reasonable level of statistical significance, the hypothesis that the coefficient estimates for models (2) and (3) in Table A.11 and models (2) and (3) in Table A.6 are all equal to one another. By implication, neither the correction for selection into employment nor the inclusion of the numeracy measure changes the earnings returns es-

timates for the numeracy sample in a statistically significant manner. The correction procedure implemented thus also fails to account for any of the statistically significant differences between the earnings returns estimates for the full sample and those for the numeracy sample.

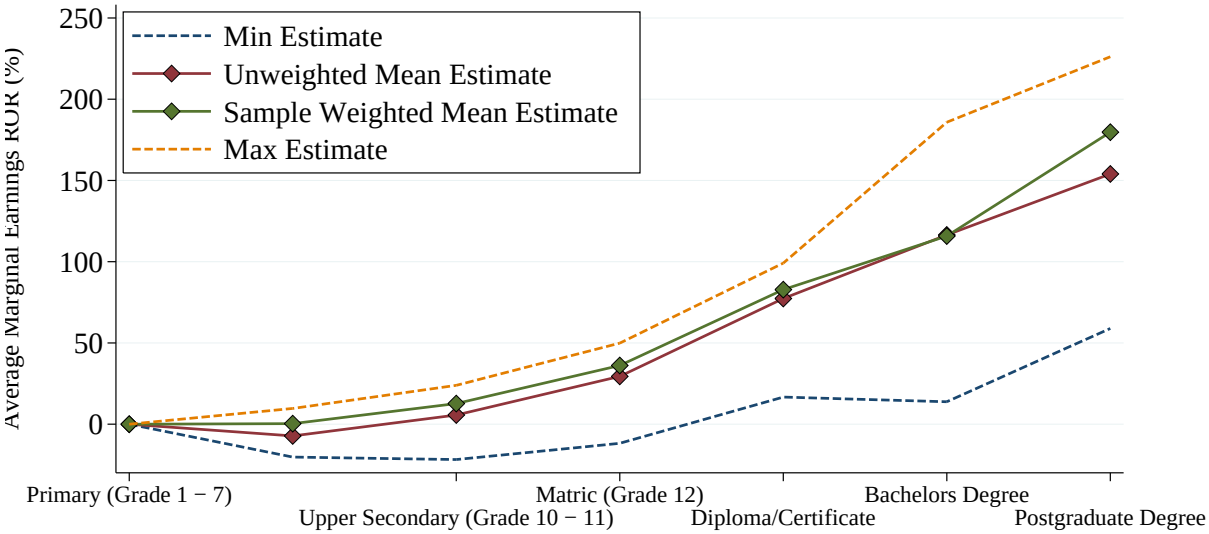
In contrast to the results for the uncorrected models in Table A.6, the coefficient on the numeracy score variable is not statistically significant in the employment-corrected models, even when the educational attainment dummy variables are excluded from the estimation in column (4). A similar result to the one observed for the employment-corrected earnings returns estimates in the school quality sample is seen in model (4) whereby the estimate of the cross-equation error correlation coefficient, $\rho_{Employment}$, is found to be large and statistically significant at the 1% level once certain variables are omitted from the second-stage earnings outcome estimation. As explained above, this result reveals the sensitivity of the Heckman (1979) MLE approach to changes in the model specification. This same result obtains, though to a lesser extent, for the numeracy-corrected earnings returns estimates in column (7) of Table A.11. Here, the estimated correlation between the residuals for the second-stage earnings outcome equation and the residuals for the first-stage numeracy test participation equation, $\rho_{Numeracy}$, is statistically significant at the 10% level. However, for models (5) and (6), $\rho_{Numeracy}$ is not statistically significant from zero, indicating that the numeracy test participation decision is not correlated with the earnings outcome in ways that have not already been controlled for by the covariates included in the second-stage estimation.

To summarize the earnings returns to educational attainment results found in this study, Figure 6.8 plots the overall minimum, unweighted mean, estimation-sample weighted mean, and maximum estimates of the average marginal earnings returns to educational attainment as calculated over eight of the corrected and uncorrected regressions presented in this Chapter.[39] Similar to the results for the employment returns to educational attainment in Figure 6.8, the differences between the minimum and maximum earnings returns estimates in the graph appear to remain fairly constant up to diploma/certificate levels of attainment, whereafter it widens somewhat owing largely to the imprecise estimation of the earnings returns to higher levels of educational attainment in the employment-corrected estimations for the school quality sample.

Based on the the estimates in Tables A.4, A.5, A.6, A.10, and A.11 the plots suggest that the earnings returns to education profile remains convex even when controlling for numeracy and school quality measures and correcting for selection into employment. In fact, insofar as the magnitudes of the returns to education coefficients are concerned, the results offer virtually no

[39] The results from the following regressions were used to calculate the summary measures: (5) in Table A.4; (3) in Table A.5; (3) in Table A.6; (1), (3) and (6) in Table A.10; and (3) and (6) from Table A.11. Collectively, these regressions are representative of the different samples, model specifications and approaches used in the etimations of earnings outcomes in South Africa. The respective estimates are assigned equal weight in the calculation of the arithmetic mean. The estimation-sample-weighted mean is calculated by weighting each coefficient estimate by the size of the estimation sample form which it was derived as a fraction of the sum of the eight respective estimation samples used in obtaining the summary measures: $Weight_i = \frac{n_{estimation_i}}{\sum_{i=1}^{8} n_{estimation_i}}$

Figure 6.10: Average Marginal Earnings Returns to Educational Attainment Summaries



NOTES: The max, unweighted mean, sample weighted mean and min estimates presented in the graph are composite summary measures that correspond to the maximum, mean, estimation-sample-weighted mean, and minimum average marginal earnings returns to educational attainment estimates calculated for the following 8 regressions: (5) in Table A.4; (3) in Table A.5; (3) in Table A.6; (1), (3) and (6) in Table A.10; and (3) and (6) from Table A.11. The average marginal rares of return (ROR) calculated for each regression represent the *ceteris paribus* discrete average percentage change in earnings relative to the base category (completed primary education): $\frac{\partial Pr(Earnings)}{(Educaton_j - Education_{base})} = (X\beta|Educ_j) - (X\beta|Educ_{base}) \ \forall j \neq base$

concrete evidence to suggest that the extent of the ability and school quality biases in the estimates of the labour market returns to educational attainment are anything other than negligible.

Chapter 7

Conclusion: Results, Findings, Caveats, and Implications

This study has attempted to provide a comprehensive assessment of the extent to which educational attainment, school quality and numeric competency improve individuals' employment and earnings prospects in the South African labour market. To provide the necessary background for the empirical analysis, an overview of the literature on the employment and earnings returns to education, school quality and numeracy in the international and local literatures was provided along with a theoretical discussion of the origins and dimensions of the human capital concept and the reasons for the widely-observed positive associations between labour market outcomes and investments in different human capital components. The 2008 National Income Dynamics Study was introduced as a suitable candidate for the empirical analysis on the basis that it would allow for measures of educational attainment, school quality and numeracy to be linked to individual labour market outcomes. The specific features of the dataset were discussed and the implications of the systematic patterns of missing observations on the NIDS numeracy and school quality scores investigated.

Due to the potential for omitted variable bias in the labour market returns to educational attainment and the potential for sample selection bias due to the selectivity of the labour force participant, employment, school quality and numeracy samples, a formal representation of the consequences of omitted variable and sample selection bias was given following an overview of the evidence on omitted variable and sample selection bias from the literature on labour market returns estimations. Because of the specific features of the NIDS data and the objectives to be pursued in the multivariate analysis it was decided that proxy varaibles should be used to correct for omitted variable bias and that the Heckman Maximum likelihood should be implemented to

correct for each of the four potential sources of sample selection bias. Lastly, the states and sociodemographic correlates of human capital endowments and labour market outcomes in South Africa were assessed and some descriptive analyses performed in order to provide the necessary context for the subsequent regression estimations.

The results from the multivariate analysis in the preceding chapter provides some insights into the nature and magnitudes of the complex relationships that exist between educational attainment, school quality, numeracy, and employment and earnings outcomes in the South African labour market. Moreover, the results reveal that the magnitudes and statistical significance of the marginal employment and earnings returns to educational attainment estimates vary substantially depending on the model specification and estimation sample used. This chapter summarizes some of the main results and findings from the multivariate analysis in terms of the six estimation objectives outline in Figure 6.2, discusses a few important caveats to those findings, and finally concludes on the implications of this study.

7.1 Main Results and Findings

The results from the standard and sample selection-corrected *probit* estimations reveal that the employment benefits of investments in education in South Africa are only significant at higher levels of educational attainment and that the structure of the employment returns to education is at least partly convex. While the employment returns to educational attainment are statistically negligible prior to the completion of Matric, the probability of being employed rises at an increasing rate as individuals attain tertiary education, remaining constant only between bachelors and post-graduate degree attainment levels. This convexity holds irrespective of the model specification, estimation procedure, and estimation sample used. However, the extent of the convexity and the magnitudes of the estimates are shown to be highly sensitive to the sample used, even when attempting to correct for the various potential instances of sample selection using the Heckman ML procedure.

Given the relatively small number of NIDS respondents with post-secondary education levels and the fact that any selection effects are likely to be exacerbated in small samples, the resultant variation in the estimates of the employment returns to education are found to be greater at higher levels of educational attainment. While individuals with post-secondary diplomas or certificates are on average estimated to receive an 8 to 18 percentage point premium over those with only completed primary schooling in terms of the probability of being employed, the corresponding figures for individuals with bachelors and post-graduate degrees are 18 to 52 and 4 to 43 percentage points respectively, depending on the sample used in the estimation. Because of the substantial differences between the minimum and maximum estimates of these employment

returns to higher levels of educational attainment, the estimation-sample-weighted mean marginal employment returns to education are calculated from the various regression results. These figures suggest that, relative to individuals with primary schooling only, a Matric qualification raises the probability of being employed by 8 percentage points, a post-secondary diploma or certificate by 12 percentage points, and a bachelors or post-graduate degree by approximately 23 percentage points on average, *ceteris paribus.* Although not directly comparable, the structure of these returns appears to be broadly consistent with that found by Oosthuizen (2006, p. 56).

In contrast to the findings from other studies cited in Section 2.2.2, the results from the multivariate analysis suggest that school quality, as measured by the NIDS school quality score, has a negligible impact on the probability of being employed once other observable employment-relevant factors have been controlled for. A similar result holds for the effect of numeracy which is found either to be statistically insignificant or to make a limited contribution to the probability of being employed in South Africa. A one standard deviation increase in numeric competency, as measured by the the NIDS numeracy score variable, is estimated to increase the probability of being employed by no more than 3 percentage points on average, *ceteris paribus*. Moreover, the magnitudes of the employment returns to education and the convex structure of those returns are found to be invariant to the inclusion of the NIDS numeracy and school quality measures. This either suggests that estimates of the employment returns to educational attainment in South Africa are not subject to omitted variable bias in terms of the omission of measures of numeric ability and school quality or simply that the NIDS numeracy and school quality scores fail to adequately proxy for such measures. In light of the existing evidence on the importance of race for the probability of being employed in the South African labour market, it is surprising that the employment returns to race are found to be statistically negligible once age, gender, marital status, household head status, physical health, emotional well-being and educational attainment levels have been controlled for.

The variation in the estimates of the employment returns to various sociodemographic and human capital indicators in this study appears to be driven almost exclusively by the properties of the different samples used in the estimations. While the descriptive analysis provides preliminary indications that both the NIDS school quality and numeracy measures may be subject to selection on unobservables that are relevant for employment outcomes, this hypothesis is not supported by the results of the multivariate analysis.[1] Consequently, it is perhaps not surprising that the attempts to correct for selection into the school quality and numeracy samples using the Heckman ML procedure produces estimates which are not statistically different from those obtained in the uncorrected estimations. However, it is somewhat surprising that the results show that selection into labour force participation does not have a statistically significant impact on

[1] As mentioned above, the finding that some unobserved correlates of the probability of employment may also be determinants of selection into the numeracy sample is largely circumstantial.

the estimates of employment returns in South Africa once common observable correlates of LFP and employment have been accounted for. These findings can be interpreted as suggesting that neither selection into LFP nor selection into the school quality or numeracy samples leads to sample selection bias in the estimates of the employment returns in the South African labour market.

Evaluated in terms of the objectives outlined in Figure 6.2, the results from the OLS and Heckman ML estimations of the earnings returns to educational attainment, school quality and numeracy are broadly similar to those discussed for the employment returns estimates above. While the convexity in the structure of the estimated earnings returns to educational attainment in South Africa appears to be stronger and more persistent than the convexity in the estimated employment returns, it is also found to be far more sensitive to the model specification and estimation sample used. The extent of the divergence between the minimum and maximum estimates of the earnings returns to educational attainment as one moves along the attainment distribution is also much larger than in the case of the employment returns. This can clearly be observed in Figure 6.10 which shows a staggering 170 percentage point difference between the minimum and maximum marginal earnings returns estimates to bachelors and post-graduate attainment levels compared to a difference of about 40 percentage points for the upper-secondary attainment level.

The substantial divergence in the estimated average marginal returns to educational attainment accrues from the significant differences in the educational attainment distributions in the various earnings function estimation samples. The estimation-sample-weighted mean marginal earnings returns to education estimates suggest that, on average, completion of upper secondary schooling is associated with 13% higher monthly labour market earnings than completion of only primary education, *ceteris paribus*. The corresponding marginal earnings returns estimates are considerably higher for Matric (35%), post-secondary diploma or certificate (83%), bachelors degree (115%) and post-graduate degree (180%) attainment levels.

In contrast to the results for the analysis of employment outcomes, school quality is found to have a significant impact on labour market earnings when race is not accounted for. A comparison of the estimated unexplained racial earnings gaps and the average marginal school quality effects for the different race groups not only shows that differences in school quality matter for labour market earnings in South Africa but also that part of the previously-unexplained racial differentials in earnings outcomes, often perceived to be the result of labour market discrimination, may be explained by racial differentials in the quality of schooling in South Africa. However, due to substantially selectivity in and miss-measurement of school quality in the NIDS data, the inclusion of the school quality measure in the earnings functions is found to have a statistically negligible impact on the coefficients on the race dummy variables. These results suggest the possibility that the racial dummies which are commonly included in earnings functions for South Africa may, to a large degree, actually stand proxy for inter-racial school quality

differences. The present study presents a potentially extreme case of this where it appears as though race may be a better proxy for school quality than the NIDS school quality measure used in the analysis.

While school quality is found to have a significant impact on labour market earnings in the absence of controls for race, the estimated earnings returns to numeracy are statistically negligible once educational attainment levels are accounted for. Similarly, the inclusion of the NIDS school quality and numeracy measures in the earnings functions has no statistically significant impacts on the magnitudes of or the convexity of the estimated marginal earnings returns to educational attainment. This can again be interpreted as suggesting either that the estimates of the earnings returns to educational attainment in South Africa are not subject to omitted variable bias due to the exclusion of numeracy and school quality measures or that the NIDS variables used to proxy for such measures in this study fail to adequately do so.

The substantial variation in the various estimated earnings returns in this study may be attributed almost entirely to the underlying differences in the various estimation samples. Based on the preliminary findings from the descriptive analysis, attempts were again made to correct for any potential bias in the estimates of the earnings returns due to endogenous selection into the numeracy, school quality, and employment samples. However, the small estimation sample sizes compromise the precision with which the coefficient and model parameters are estimated when implementing the Heckman ML procedure. Consequently, the estimates of the correlation between the first and second-stage selection and outcome equation residuals are statistically significant and large for some estimations and statistically insignificant for others. However, the estimates obtained using the Heckman ML procedure are found to be no different from those obtained under normal OLS, even when the model statistics suggest that it is necessary to account for selection bias. Again, the results from the estimated models suggest that selection into employment does not have a statistically significant impact on the estimates of the earnings returns in South Africa once common observable correlates of employment and labour market earnings outcomes have been controlled for. It must therefore again be concluded that the estimates of the earnings returns to educational attainment in South Africa may not be subject to selection bias emanating from selection into employment, or selection into the numeracy or school quality samples.

7.2 Important Caveats

The validity of the findings reported above and elsewhere in this study are conditional on a number of crucial assumptions, many of which may be contested on several grounds. As such it is necessary to consider some of the caveats to the findings from the empirical analysis before

concluding on the implications of this study. These caveats relate to a number of theoretical considerations and practical issues pertaining to the measurement of the NIDS numeracy and school quality variables.

7.2.1 Theoretical issues

The empirical analysis presented above abstracts entirely from the consideration and modelling of the costs of human capital investments and how those costs influence individuals' human capital investment decisions. Instead, it is assumed that individuals' human capital stock endowments at any point in time can be taken as given. This assumption simplifies the analysis of the association between human capital and labour market outcomes in South Africa, but may be criticized on the basis that one cannot accurately quantify the returns to investments in human capital and compare them to the returns on other investments without explicitly accounting for the differences in the expected costs of, and the expected payoffs to, those investments. By extension, it would be imprudent to propagate the profitability of investments in human capital without understanding how the returns on those investments relate to their costs. In order to adequately incorporate such cost considerations in multivariate analyses, it is necessary to impose additional theoretical structure on the estimation methodology. However, such a structural approach is beyond the scope of this particular study.[2]

The absence of an explicit theoretical model governing the structure of the labour market returns to human capital estimations also has important implications for the causal interpretation of the regression estimates. While it is common for researchers to try and draw causal inferences from non-structural models (such as those used in this study) based on *a priori* expectations, the parameter estimates from such models can only be interpreted as partial correlations. As such, it cannot be claimed that the estimated marginal employment and earnings "returns" to educational attainment, school quality, and numeracy in this study are accurate representations of the causal impacts that these three measures of human capital have on labour market outcomes in South Africa. Instead, the estimated coefficient on any particular variable of interest simply measures the degree of the association which it shares with the outcome variable once the effects of other observables have been partialled out.

While dealing with potential omitted variable and sample selection biases in the estimation of the South African labour market returns to human capital is emphasized throughout this study, the analysis does not account for the fact that the human capital variables considered may be endogenous to the labour market outcome estimations they are meant to explain. As discussed in Chapters 1 and 4, it is likely that some of the unobservable factors that influence the probability of employment and earnings capacity also influence levels of educational attainment, school

[2] See Burger (2011), for example, for a structural estimation of the labour market earnings returns to education in South Africa.

quality, and numeric competency in South Africa.[3] In theory, instrumental variable (IV) techniques could be used to try and correct for any potential resultant endogeneity biases, provided that valid instruments are available in the data (Chen and Hamori, 2009, p. 145). However, the lack of sufficiently adequate IVs in the NIDS 2008 dataset make such an approach infeasible.[4] Consequently, the possibility that the estimates of the employment and earnings returns to education, school quality, and numeracy will be subject to endogeneity bias cannot be ruled out.

7.2.2 Data Issues and Practical Considerations

While NIDS 2008 is one of the first datasets to directly allow for numeracy and school quality information to be linked to individual labour market outcomes in South Africa, the substantial number of missing observations and the highly selective pattern of observability on the NIDS school quality and numeracy score variables limit the scope for obtaining unbiased estimates of the labour market returns to numeracy and school quality in South Africa with any reasonable degree of precision. The adverse implications of these deficiencies are evident in the volatility displayed by the magnitudes and levels of statistical significance of the multivariate analysis results. In addition, the finding that neither numeric competency nor school quality has a robust and statistically significant impact on employment or earnings prospects (when controlling for other factors) contradicts the existing theoretical literature and empirical evidence on the labour market returns to these measures of human capital in South Africa.[5] In order to reconcile the findings in this study with those found in the literature, the following issues pertaining to the data thus need to be taken into account.

The finding that the NIDS numeracy and school quality variables have statistically negligible impacts on the NIDS employment and earnings outcome variables does not necessarily imply that numeracy and school quality have negligible impacts on employment and earnings outcomes in South Africa, On the contrary, this result is likely to derive mainly from the deficiencies in the NIDS numeracy and school quality variables. The results from the descriptive analysis suggest that both the NIDS numeracy and school quality score variables fail to adequately capture the lower tails of the numeracy and school quality distributions for the South African working-age population. As such, these variables are likely to be upward-biased indicators of the average levels of numeric competency and school quality in the country. In

[3] See Dougherty (2003) and Charette and Meng (1998) for a discussion on the endogeneity of educational attainment and numeracy and Case and Yogo (1999, p. 2) for a discussion on the endogenous nature of school quality in South Africa.

[4] See Du Rand *et al.* (2011) for a discussion on the use of IVs when estimating the earnings returns to numeracy using the NIDS 2008 data.

[5] See Moll (1992, 1998), Pillay (1994), Case and Yogo (1999), Chamberlain and Van der Berg (2002), Burger and Van der Berg (2011), Moses (2011), and Yamauchi (2004, 2011).

theory and with all else being held constant, this should imply that the use of these measures in labour market returns estimations will produce downward-biased estimates of the marginal returns to numeracy and school quality in South Africa. The greater the extent to which the NIDS numeracy and school quality measures fail to capture the lower regions of the true numeracy and school quality distributions, the greater the expected attenuation in the magnitude of the estimated employment and earnings returns to numeracy and school quality.

In addition to the expected downward-bias in the estimates of the labour market returns to school quality and numeracy, left-censoring of the NIDS numeracy and school quality variables suggests that they understate the extent of the variation in numeric competency and school quality in South Africa. This effect is exacerbated by conversion of the standardized school quality score into discrete values which results in a further loss of variation in the NIDS school quality variable. The limited variation in these two human capital measures coupled with the small numeracy and school quality sample sizes will have adverse implications for the achievement of the estimation objectives if it implies that the estimations samples are homogenised to such an extent that the parameters of interest can no longer be estimated with any reasonable level of precision. This could partly explain why many of the estimated coefficients in the sample-selection-corrected models are found to be statistically insignificant despite being relatively large in magnitude.

7.3 Conclusions and Implications

In order to improve the labour market prospects faced by South Africans and redress the persistent socio-economic inequalities in South African society, it is necessary to understand the role that human capital plays in determining labour market outcomes in South Africa. Moreover, for policy interventions to be effective it is necessary to understand which specific components of human capital are most important for achieving success in the South African labour market. This study has endeavoured to provide a comprehensive analysis of the employment and earnings benefits to three such components of human capital: educational attainment, school quality and numeracy. While the results from the empirical analysis are plagued by data deficiencies and fail to provide definitive answers to many of the objectives that are pursued, a number of important qualitative conclusions can be drawn from the study's findings.

While there are some indications of significant employment and earnings returns to completed secondary education in South Africa, the labour market benefits of educational attainment appear to manifest predominantly at post-secondary levels of attainment. The labour market returns to tertiary qualifications are shown to be large and, with respect to earnings, exhibit considerable convexity even when differences in English language proficiency, computer literacy,

emotional well-being, physical health, numeracy and school quality are accounted for. As such, investments in tertiary education appear to be the most effective way, on average and with all else being held constant, of improving one's employment and earnings prospects in the South African labour market. However, such a conclusion is problematic given that the generally poor quality of primary and secondary schooling in South Africa makes tertiary education inaccessible to the vast majority of the South African population. Despite the fact that the multivariate analysis offers no convincing proof that school quality is positively and significantly associated with labour market outcomes in South Africa, there is thus clearly a role for government to play in improving the quality of education received in primary and secondary school, if only to improve individuals' capacities to accede to and benefit from higher levels of attainment.

The empirical analysis shows that while Whites have no unexplained advantage over Coloureds, Indians, or Blacks in terms of the probability of being employed, they are predicted to receive higher earnings, on average, than Blacks and Coloureds even once inter-racial differences in a wide range of observables have been taken into account. However, there is some circumstantial evidence that a significant part of the racial differentials in labour market earnings in South Africa may be explained by racial differentials in the quality of education attained. These findings are consistent with those found in other studies (see, for example, Du Rand *et al.* (2011) and Burger and Van der Berg (2011)) and suggest that much of what is often perceived as labour market discrimination may be rooted in the persistent inequalities in South Africa's formal education system. This reinforces the need for interventions aimed at improving the quality of education, particularly in historically-Black schools.

In contrast to the findings by Du Rand *et al.* (2011), the multivariate analysis in this study suggests that the labour market returns to numeric competency in South Africa are trivial in relation to the returns on investments in other components of human capital. However, the findings pertaining to the South African labour market benefits of numeracy and school quality are undermined by the deficiencies in the variables used to proxy for these two measures of human capital. Both the NIDS numeracy and school quality score variables are subject to significant and selective non-observability which leads to highly unstable estimates of the marginal earnings and employment returns to school quality and numeracy.

Finally, the differences between the results from the various estimations in the multivariate analysis appear to be largely unaffected by the attempts to correct for instances of endogenous selection using the Heckman ML procedure and instead are shown to be a function of the differences in the model specifications and the underlying features of the various estimations samples. This finding suggests that researchers should be weary of drawing strong conclusions from single sets of point estimates that are based on data which fail to accurately represent the population of interest. Consequently, some form of sensitivity analysis is warranted whenever the imperfections in the data threaten to undermine the robustness of one's findings. More importantly, however, the results show that the scope for overcoming data deficiencies by using

standard parametric estimation techniques may be limited when the extent of those deficiencies are severe. In order to understand the true extent to which measures such as numeracy and school quality influence labour market outcomes in South Africa, it is thus essential that accurately-measured and representative data on human capital measures and labour market outcomes are collected. Unless the quality of data which is currently available improves significantly, assessments of the labour market returns to different human capital components in South Africa may remain little more than a technical exercise and continue to produce results which are largely uninformative for the purposes of policy.

Bibliography

AKOOJEE, S., MCGRATH, S. and VISSER, M. (2008). Further Education and Training Colleges. In A. Kraak and K. Press (Eds.) *Human Resources Development Review 2008: Education, Employment and Skills in South Africa*, chap. Further education and training colleges, Cape Town: HSRC Press.

ANDERSON, K. G., CASE, A. and LAM, D. (2001). Causes and Consequences of Schooling Outcomes in South Africa: Evidence from Survey Data. *Social Dynamics* **27**(1), pp. 37 – 59.
[online] available from http://casr.ou.edu/pubs/KA_Causes__Consequences.pdf

ARABSEIBANI, G. R. and REES, H. (1998). On the Weak vs Strong Version of the Screening Hypothesis: A Re-Examination of the P-Test for the U.K. *Economics of Education Review* **17**(2), pp. 189–192.
[online] available from http://dx.doi.org/doi:10.1016/0165-1765(79)90232-5

ARIAS, O., HALLOCK, K. F. and SOSA-ESCUDERO, W. (2001). Individual Heterogeneity in the Returns to Schooling: Instrumental Variables Quantile Regression Using Twins Data. *Empirical Economics* **26**, pp. 7–40.
[online] available from http://dx.doi.org/10.1007/s001810000053

ASHENFELTER, O. and KRUEGER, A. (1994). Estimates of the Economic Return to Schooling from a New Sample of Twins. *The American Economic Review* **84**(5), pp. 1157–1173.
[online] available from http://www.jstor.org/stable/2117766

ASHENFELTER, O. and ROUSE, C. (1998). Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins. *The Quarterly Journal of Economics* **113**(1), pp. 253–284.
[online] available from http://www.jstor.org/stable/2586991

AXINN, W. G. and PEARCE, L. D. (2006). *Mixed method data collection strategies*. New York: Cambridge University Press.

AZAM, J.-P. and ROSPABÉ, S. (2007). Trade Unions vs. Statistical Discrimination: Theory and Application to Post-apartheid South Africa. *Journal of Development Economics* **84**, pp.

417–444.
[online] available from http://dx.doi.org/doi:10.1016/j.jdeveco.2005.12.005

BANERJEE, A., GALIANI, S., LEVINSOHN, J., MCLAREN, Z. and WOOLARD, I. (2006). Why Has Unemployment Risen in the New South Africa? IPC Working Paper 35, International Policy Center (IPC), Gerald R. Ford School of Public Policy, University of Michigan.
[online] available from http://www.nber.org/papers/w13167

BECKER, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *The Journal of Political Economy* **70**(5), pp. 9–49.
[online] available from http://www.jstor.org/stable/1829103

——— (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: Columbia University Press.

——— (1992). Human Capital and the Economy. *Proceedings of the American Philosophical Society* **136**(1), pp. 85–92.
[online] available from http://www.jstor.org/stable/986801

BEHRMAN, J. R. and ROSENZWEIG, M. R. (1999). "Ability" Biases in Schooling Returns and Twins: A Test and New Estimates. *Economics of Education Review* **18**, pp. 159–167.
[online] available from http://dx.doi.org/doi:10.1016/S0272-7757(98)00033-8

BEKKER, P. A. and WANSBEEK, T. J. (1996). Proxies Versus Omitted Variables in Regression Analysis. *Linear Algebra and its Applications* **237-238**, pp. 301 – 312.
[online] available from http://www.sciencedirect.com/science/article/pii/0024379594005915

BETRAND, M., MULLAINATHAN, S. and MILLER, D. (2003). Public Policy and Extended Families: Evidence from Pensions in South Africa. *The World Bank Economic Review* **17**(1), pp. 27–50.
[online] available from http://dx.doi.org/DOI:10.1093/wber/lhg014

BHORAT, H. and MCCORD, A. (2003). *Human Resources Development Review. Education, Employment and Skills in South Africa*, chap. Employment and Labour Market Trends, pp. 112–141. Research Programme on Human Resource Development, HSRC. Paarl: Paarl Print.

BLACKBURN, M. L. and NEUMARK, D. (1993). Omitted-Ability Bias and the Increase in the Return to Schooling. *Journal of Labor Economics,* **11**(3), pp. 521 – 544.
[online] available from http://www.jstor.org/stable/2535084

BLAUG, M. (1972). The Correlation between Education and Earnings: What Does It Signify? *Higher Education* **1**(1), pp. 53–76.
[online] available from http://www.jstor.org/stable/3445959

——— (1976). The Empirical Status of Human Capital Theory: A Slightly Jaundiced Survey. *Journal of Economic Literature* **14**(3), pp. 827–855.
[online] available from http://www.jstor.org/stable/2722630

BLUNDELL, R., DEARDEN, L., MEGHIR, C. and SIANESI, B. (1999). Human Capital Investment: The Returns from Education and Training to the Individual, the Firm and the Economy. *Fiscal Studies* **20**(1), pp. 1–23.
[online] available from http://dx.doi.org/DOI:10.1111/j.1475-5890.1999.tb00001.x

BONJOUR, D., CHERKAS, L. F., HASKEL, J. E., HAWKES, D. D. and SPECTOR, T. D. (2003). Returns to Education: Evidence from U.K. Twins. *The American Economic Review* **93**(5), pp. 1799–1812.
[online] available from http://www.jstor.org/stable/3132153

BRADBURN, N. M. (1978). Respondent Burden. In L. Reeder (Ed.) *Health Survey Research Methods: Second Biennial Conference*, vol. Williamsburg, Va., Washington, D.C.: US Government Printing Office.

BRANSON, N., LEIBBRANDT, M. and ZUZE, T. L. (2009). *Determining the Scope of the Problem and Developing a Capacity-Building Model*, chap. What are the returns of tertiary education and who benefits?, pp. 45–60. Centre for Higher Education Transformation (CHET)/ African Minds, 1st edn.

BROWN, S. and SESSIONS, J. G. (1998). Education, Employment Status and Earnings: A Comparative Test of the Strong Screening Hypothesis. *Scottish Journal of Political Economy* **45**(5), pp. 586–591.
[online] available from http://dx.doi.org/DOI:10.1111/1467-9485.00113

——— (2006). Evidence on the Relationship between Firm-based Screening and the Returns to Education. *Economics of Education Review* **25**, pp. 495–509.
[online] available from http://dx.doi.org/doi:10.1016/j.econedurev.2005.05.007

BURGER, C. (2008). Sample Selection Bias and the South African Wage Function. Stellenbosch Economic Working Paper 18/08, Department of Economics and the Bereau for Economic Research at the University of Stellenbosch.
[online] available from http://www.ekon.sun.ac.za/wpapers/2008

BURGER, C. and VAN DER BERG, S. (2011). Modelling Cognitive Skills, Ability and School Quality to Explain Labour Market Earnings Differentials. Stellenbosch Economic Working Paper 08/11, Department of Economics and the Bereau for Economic Research at the University of Stellenbosch.
[online] available from http://www.ekon.sun.ac.za/wpapers/2011

BURGER, R. (2011). *Estimating the shape of the South African schooling-earnings profile*. Ph.D. thesis, Oriel College, Oxford University.

BURGER, R. and VON FINTEL, D. (2009). Determining the Causes of the Rising South African Unemployment Rate: An Age, Period and Generational Analysis. Stellenbosch Economic Working Paper 24/2009, Department of Economics and the Bereau for Economic Research at the University of Stellenbosch.
[online] available from http://www.ekon.sun.ac.za/wpapers/2009

CAMERON, A. C. and TRIVEDI, P. K. (2009). *Microeconometrics Using Stata*. College Station: Stata Press.

CARD, D. and KRUEGER, A. B. (1996). Labor Market Effects of School Quality: Theory and Evidence. Working Paper 5450, National Bureau of Economic Research.
[online] available from http://www.nber.org/papers/w5450

CASALE, D. and POSEL, D. (2010). Unions and the Gender Wage Gap in South Africa. *Journal of African Economies* **20**(2), pp. 27–59.
[online] available from http://dx.doi.org/doi:10.1093/jae/ejq029

CASE, A. and YOGO, M. (1999). Does School Quality Matter? Returns to Education and the Characteristics of Schools in South Africa. NBER Working Papers 7399, National Bureau of Economic Research, Inc.
[online] available from http://econpapers.repec.org/RePEc:nbr:nberwo:7399

CASTAGNETTI, C., CHELLI, F. and ROSTI, L. (2005). Educational Performance as Signalling Device: Evidence from Italy. *Economics Bulletin* **9**(4), pp. 1–7.
[online] available from http://ideas.repec.org/a/ebl/ecbull/v9y2005i4p1-7.html

CHAMBERLAIN, D. and VAN DER BERG, S. (2002). Earnings Functions, Labour Market Discrimination and Quality of Education in South Africa. Working Papers 02/2002, Stellenbosch University, Department of Economics.
[online] available from http://ideas.repec.org/p/sza/wpaper/wpapers2.html

CHARETTE, M. F. and MENG, R. (1998). The Determinants of Literacy and Numeracy, and the Effect of Literacy and Numeracy on Labour Market Outcomes. *The Canadian Journal of Economics* **31**(3), pp. 495 – 517.
[online] available from http://www.jstor.org/stable/136200

CHEN, G. and HAMORI, S. (2009). Economic Returns to Schooling in Urban China: OLS and the Instrumental Variables Approach. *China Economic Review* **20**, pp. 143–152.
[online] available from http://dx.doi.org/doi:10.1016/j.chieco.2009.01.003

CHEVALIER, A., GIBBONS, S., THORPE, A., SNELL, M. and HOSKINS, S. (2009). Students' Academic Self-Perception. *Economics of Education Review* **28**(6), pp. 716–727.
[online] available from http://dx.doi.org/doi:10.1016/j.econedurev.2009.06.007

CLARK, A. (2000). Signalling and Screening in a Transition Economy Three Empirical Models Applied to Russia. CERT Discusson Paper 0003, Centre for Economic Reform and Transformation, Heriot Watt University.
[online] available from http://ideas.repec.org/p/hwe/certdp/0003.html

COBBEN, F. (2009). *Nonresponse in Sample Surveys: Methods for Analysis and Adjustment*. The Haguge: Statistics Netherlands.

COLCLOUGH, C., KINGDON, G. and PATRINOS, H. A. (2008). The Pattern of Returns to Education and its Implications. RECOUP Policy Brief Number 4, Research Consortium on Educational Outcomes & Poverty, Cambridge, UK.
[online] available from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-69182

COUPER, M. P. and SINGER, E. (2009). The Role of Numeracy in Informed Consent for Surveys. *Journal of Empirical Research on Human Research Ethics* **4**(4), pp. 17–26.
[online] available from http://dx.doi.org/doi:10.1525/jer.2009.4.4.17

DANIELS, R. (2007). Skills Shortages in South Africa: A Literature Review. Working Papers 9697, University of Cape Town, Development Policy Research Unit.
[online] available from http://ideas.repec.org/p/ctw/wpaper/9697.html

DAVIDSON, R. and MACKINNON, J. G. (2004). *Econometric Theory and Methods*. New York: Oxford University Press.

DEPARTMENT OF LABOUR (2003). State of Skills in South Africa 2003. Tech. rep., Skills Development Planning Unit, Department of Labour.

DIAS, R. and POSEL, D. (2007). Unemployment, Education and Skills Constraints in Post-Apartheid South Africa. DPRU Working Paper 07/120, Development Policy Research Unit.
[online] available from http://ssrn.com/abstract=982046

DOLTON, P. J. and SILLES, M. A. (2008). The Effects of Over-education on Earnings in the Graduate Labour Market. *Economics of Education Review* **27**(2), pp. 125–139.
[online] available from http://dx.doi.org/DOI:10.1016/j.econedurev.2006.08.008

DOUGHERTY, C. (2003). Numeracy, Literacy and Earnings: Evidence from the National Longitudinal Survey of Youth. *Economics of Education Review* **22**(2003), pp. 511 – 521.
[online] available from http://dx.doi.org/doi:10.1016/S0272-7757(03)00040-2

DU RAND, G., VAN BROEKHUIZEN, H. and VON FINTEL, D. (2010). Who Responds to Voluntary Cognitive Tests in Household Surveys? The Role of Labour Market Status, Respondent Confidence, Motivation and a Culture of Learning in South Africa. Stellenbosch Economic Working Paper 27/10, Department of Economics and the Bereau for Economic Research at the University of Stellenbosch.
[online] available from http://ideas.repec.org/p/sza/wpaper/wpapers126.html

——— (2011). Numeric Competence, Confidence and School Quality in the South African Wage Function: Towards Understanding Pre-labour Market Discrimination. Stellenbosch Economic Working Paper 12/2011, Department of Economics and the Bereau for Economic Research at the University of Stellenbosch.
[online] available from http://ideas.repec.org/p/sza/wpaper/wpapers140.html

FERTIG, M. and SCHURER, S. (2007). Labour Market Outcomes of Immigrants in Germany: The Importance of Heterogeneity and Attrition Bias. IZA Discussion Paper 2915, Institute for the Study of Labor.
[online] available from http://ideas.repec.org/p/rwi/repape/0020.html

GILLEY, O. W. and LEONE, R. P. (1991). A Two-Stage Imputation Procedure for Item Nonresponse in Surveys. *Journal of Business Research* **22**, pp. 281 – 291.
[online] available from http://dx.doi.org/doi:10.1016/0148-2963(91)90035-V

GLAZERMAN, S., SCHOCHET, P. Z. and SCHOCHET, P. Z. (2000). National Job Corps Study: The Impacts of Job Corps on Participants' Literacy Skills. Mpr report, U.S. Department of Labor, Employment and Training Administration Office of Policy and Research Report. Mathematica Policy Research, Inc.
[online] available from http://wdr.doleta.gov/opr/fulltext/00-jc_literacy.pdf

GREENE, W. H. (2002). *Econometric Analysis*. New Jersey: Prentic Hall, 5th edn.

GRIFFIN, P., LEIBBRANDT, M., PAVLOVIC, M. and ZUZE, T. L. (2010). Numeric Literacy in South Africa: Report on the NIDS Wave 1 Numeracy Test. N.I.D.S. Technical Paper 5, National Income and Dynamics Study.

GUSTAFSSON, M. (2011). The When and How of Leaving School: The Policy Implications of New Evidence on Secondary Schooling in South Africa. Stellenbosch Economic Working Papers: 09/11, Department of Economics and The Bereau for Economic Research at the University of Stellenbosch.
[online] available from www.ekon.sun.ac.za/wpapers/2011/wp092011/wp-09-2011.pdf

HANUSHEK, E. A. (2005). The Economics of School Quality. *German Economic Review* **6**(3), pp. 269–286.
[online] available from http://dx.doi.org/DOI:10.1111/j.1468-0475.2005.00132.x

HANUSHEK, E. A., LAVY, V. and HITOMI, K. (2008). Do Students Care about School Quality? Determinants of Dropout Behavior in developing Countries. *Journal of Human Capital* **2**(1), pp. 69–105.
[online] available from http://www.jstor.org/stable/10.1086/529446

HANUSHEK, E. A. and WOESSMANN, L. (2010). Sample Selectivity and the Validity of International Student Achievement Tests in Economic Research. NBER Working Paper 15867, National Bureau of Economic Research (NBER).
[online] available from http://www.nber.org/papers/w15867

HARMON, C., OOSTERBEEK, H. and WALKER, I. (2003). The Returns to Education: Microeconomics. *Journal of Economic Surveys* **17**(2), pp. 115–155.
[online] available from http://dx.doi.org/DOI:10.1111/1467-6419.00191

HECKMAN, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* **47**(1), pp. 153–161.
[online] available from http://www.jstor.org/stable/1912352

ISACSSON, G. (2004). Estimating the Economic Return to Educational Levels Using Data on Twins. *Journal of Applied Econometrics* **19**(1), pp. 99–119.
[online] available from http://www.jstor.org/stable/25146269

JOHNSTON, J. and DINARDO, J. (1996). *Econometric Methods*. New York: McGraw-Hill, 4th edn.

KERCKHOFF, A. C., RAUDENBUSH, S. W. and GLENNIE, E. (2001). Education, Cognitive Skill, and Labor Force Outcomes. *Sociology of Education* **74**(1), pp. 1–24.
[online] available from http://www.jstor.org/stable/2673142

KESWELL, M. (2004). Education and Racial Inequality in Post Apartheid South Africa. Working Paper No. 2004-02-008, Santa Fe Institute, Santa Fe, New Mexico.

KESWELL, M. and POSWELL, L. (2002). How Important is Education for Getting Ahead in South Africa. CSSR Working Paper No. 22, Center for Social Science Research, Social Surveys Unit, University of Cape Town.
[online] available from www.cssr.uct.ac.za/publications/working-paper/2002/022

——— (2004). Returns to Education in South Africa: A Retrospective Sensitivity Analysis of the Available Evidence. *Discrimination, Income Distribution, Education* **72**(4), pp. 834–860.
[online] available from http://dx.doi.org/DOI:10.1111/j.1813-6982.2004.tb00136.x

KINGDON, G. and KNIGHT, J. (2007). Unemployment in South Africa, 1995-2003: Causes, Problems and Policies. *Journal of African Economies* **16**(5), pp. 813 – 848.
[online] available from http://dx.doi.org/doi:10.1093/jae/ejm016

KINGSTON, P. W., HUBBARD, R., LAPP, B., SCHROEDER, P. and WILSON, J. (2003). Why Education Matters. *Sociology of Education* **76**(1), pp. 53–70.
[online] available from http://www.jstor.org/stable/3090261

KNIGHT, J. and YUEH, L. Y. (2002). The Role of Social Capital in the Labour Market in China. Discussion Paper 121, Department of Economics, University of Oxford.
[online] available from http://ideas.repec.org/p/oxf/wpaper/121.html

KOCH, S. F. and NTEGE, S. S. (2006). Education Screening in South Africa's Labour Market? In *Accelerated and Shared Growth in South Africa: Determinants, Constraints and Opportunities 18 - 20 October 2006*, Development Policy Research Unit.
[online] available from www.tips.org.za/files/Education_Screening-StevenKochSimonSNtege.pdf

——— (2008). Returns to Schooling: Skill Accumulation or Information Revelation? Working Paper 2008-12, Department of Economics, University of Pretoria.
[online] available from web.up.ac.za/UserFiles/WP_2008_12.pdf

KRISHNAN, P. (1990). The Economics of Moonlighting: A Double Self-Selection Model. *The Review of Economics and Statistics* **72**(2), pp. 361–367.
[online] available from http://www.jstor.org/stable/2109729

LAM, D., LEIBBRANDT, M. and MLATSHENI, C. (2008). Education and Youth Unemployment in South Africa. SALDRU Working Paper 22, Southern Africa Labour and Development Research Unit, University of Cape Town.
[online] available from http://ideas.repec.org/p/ldr/wpaper/22.html

LEIBBRANDT, M., LEVINSOHN, J. and MCCRARY, J. (2005). Incomes in South Africa Since the Fall of Apartheid. Discussion Paper No. 536, Research Seminar in International Economics, University of Michigan.
[online] available from www.nber.org/papers/w11384

LEIBBRANDT, M., WOOLARD, I. and DE VILLIERS, L. (2009*a*). Methodology: Report on NIDS Wave 1. N.I.D.S. Technical Paper 1, National Income and Dynamics Study.

LEIBBRANDT, M., WOOLARD, I., MCEWEN, H. and KOEP, C. (2009*b*). Employment and Inequality Outcomes in South Africa. Technical report, Southern Africa Labour and Development Research Unit (SALDRU) and School of Economics, University of Cape Town.
[online] available from www.oecd.org/dataoecd/17/14/45282868.pdf

LEUNG, R., STAMPINI, M. and VENCATACHELLUM, D. (2009). Does Human Capital Protect Workers against Exogenous Shocks? South Africa in the 2008-2009 Crisis. IZA Discussion Paper 4608, Institute for the Study of Labor.
[online] available from http://ideas.repec.org/p/iza/izadps/dp4608.html

LUBOTSKY, D. and WITTENBERG, M. (2006). Interpretation of Regressions With Multiple Proxies. *The Review of Economics and Statistics* **88**(3), pp. 549–562.
[online] available from http://dx.doi.org/dx.doi.org/10.1162/rest.88.3.549

MARIOTTI, M. and MEINECKE, J. (2009). Nonparametric Bounds on Returns to Education in South Africa: Overcoming Ability and Selection bias. Working Paper 510, The Australian National University.
[online] available from http://econpapers.repec.org/RePEc:acb:cbeeco:2009-510

MCINTOSH, S. and VIGNOLES, A. (2001). Measuring and Assessing the Impact of Basic Skills on Labour Market Outcomes. *Oxford Economic Papers* **53**(3), pp. 453–481.
[online] available from http://dx.doi.org/doi:10.1093/oep/53.3.453

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

MLATSHENI, C. and ROSPABE, S. (2002). Why is Youth Unemployment so High and Unequally spread in South Africa? DPRU Working Paper 02/65, Development Policy Research Unit.
[online] available from www.commerce.uct.ac.za/Research_Units/dpru/.../PDF_Files/wp65.pdf

MOHANTY, M. S. (2001*a*). Determination of Participation Decision, Hiring Decision, and Wages in a Double Selection Framework: Male-Female Wage Differentials in the U.S. Labor Market Revisited. *Contemporary Economic Policy* **19**(2), pp. 197–212.
[online] available from http://dx.doi.org/DOI:10.1111/j.1465-7287.2001.tb00061.x

——— (2001*b*). Testing for the Specification of the Wage Equation: Double Selection Approach or Single Selection Approach. *Applied Economics Letters* **8**, pp. 525–529.
[online] available from http://dx.doi.org/DOI:10.1080/135048500011957

MOLL, P. G. (1992). Quality of Education and the Rise in returns to Schooling in South Africa, 1975-1985. *Economics of Education Review* **11**(1), pp. 1–10.
[online] available from http://dx.doi.org/10.1016/0272-7757(92)90016-V

——— (1998). Primary Schooling, Cognitive Skills and Wages in South Africa. *Economica* **65**(258), pp. 263–284.
[online] available from http://www.jstor.org/stable/2555147

MOSES, E. (2011). Quality of Education and the Labour Market: A Conceptual and Literature Overview. Stellenbosch Economic Working Paper 07/11, Department of Economics and the Bereau for Economic Research at the University of Stellenbosch.
[online] available from http://ideas.repec.org/p/sza/wpaper/wpapers135.html

MOTALA, S. (1995). Surviving the System-A Critical Appraisal of Some Conventional Wisdoms in Primary Education in South Africa. *Comparative Education,* **31**(2), pp. 161–179.
[online] available from http://www.jstor.org/stable/3099645

——— (2001). Quality and Indicators of Quality in South African Education: A Critical Appraisal. *International Journal of Educational Development* **21**, pp. 61–78.
[online] available from www.elsevier.com/locate/ijedudev

NAWATA, K. (1994). Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator. *Economic Letters* **45**, pp. 33–40.
[online] available from http://dx.doi.org/doi:10.1016/0165-1765(94)90053-1

NIDS (2009). NIDS Wave 1 Overview 2009. Technical report, National Income and Dynamics Study.

OOSTHUIZEN, M. (2006). The Post-Apartheid Labour Market: 1995 -2004. DPRU Working Paper 06/103, Development Policy Research Unit (DPRU).
[online] available from http://ideas.repec.org/p/ctw/wpaper/9618.html

OOSTHUIZEN, M. and BHORAT, H. (2004). The Post-Apartheid South African Labour Market. In *Conference on African Development and Poverty Reduction: The Macro-Micro Linkage*, Somerset West, South Afric.

——— (2006). Educational Outcomes in South Africa: A Production Function Approach. DPRU Working Paper 2006/5, Development Policy Research Unit (DPRU).
[online] available from http://hdl.handle.net/10625/33364

PANDAY, S. and ARENDS, F. (2008). School Drop-outs and Imprisoned Youths. *HSRC Review* **6**(1), pp. 4–5.
[online] available from http://www.hsrc.ac.za/HSRC_Review_Article-82.phtml

PARSONS, S. and BYNNER, J. (2005). Does Numeracy Matter More? Nrdc report, National Research and Development Centre for Adult Literacy and Numeracy, Institute of Education, University of London.
[online] available from www.nrdc.org.uk/download.asp?f=2979&e=pdf

PAUW, K., OOSTHUIZEN, M. and VAN DER WESTHUIZEN, C. (2008). Graduate Unemployment in the Face of Skills Shortages: A Labour Market Paradox. *South African Journal of Economics* **76**(1), pp. 45–57.
[online] available from http://dx.doi.org/DOI:10.1111/j.1813-6982.2008.00152.x

PETERS, E., VÄSTFJÄLL, D., SLOVIC, P., MERTZ, C., MAZZOCCO, K. and DICKERT, S. (2006). Numeracy and Decision Making. *Psychological Science* **17**(5), pp. 407 – 413.
[online] available from http://dx.doi.org/doi:10.1111/j.1467-9280.2006.01720.x

PIGOU, A. C. (1928). *A Study in Public Finance*. London: Macmillan and Co.

PILLAY, P. (1994). Quality of Schooling, Certification and Earnings in South Africa. *International Journal of Educational Development* **14**(1), pp. 13–22.
[online] available from http://dx.doi.org/doi:10.1016/0738-0593(94)90004-3

POSEL, D., A.FAIRBURN, J. and LUND, F. (2004). Labour Migration and Households: A Reconsideration of the Effects of the Social Pension on Labour Supply in South Africa. Forum paper, African Development and Poverty Reduction: The Macro-Micro Linkage, Lord Charlse Hotel, Somerset West, South Africa.
[online] available from http://www.tips.org.za/node/795

POSEL, D. and CASALE, D. (2011). Language Proficiency and Language Policy in South Africa: Findings from New Data. *International Journal of Educational Development* **32**, pp. 443–451.
[online] available from http://dx.doi.org/doi:10.1016/j.ijedudev.2010.09.003

PSACHAROPOULOS, G. (1973). *Returns to education: An international comparison*. San Francisco: Elsevier, Jossey-Bass.

PSACHAROPOULOS, G. and PATRINOS, H. A. (2002). Returns to Investment in Education: A Further Update. World Bank Policy Research Working Paper 2881, World Bank.
[online] available from http://ideas.repec.org/p/wbk/wbrwps/2881.htm

PUHANI, P. A. (2000). The Heckman Correction for Sample Selection and its Critique. *Journal of Economic Surveys* **14**(1), pp. 53–68.
[online] available from http://dx.doi.org/DOI:10.1111/1467-6419.00104

RIVERA-BATIZ, F. L. (1992). Quantitative Literacy and the Likelihood of Employment among Young Adults in the United States. *The Journal of Human Resources* **27**(2), pp. 313 – 328.
[online] available from http://www.jstor.org/stable/145737

SCHULTZ, T. W. (1961). Investment in Human Capital. *The American Economic Review* **51**(1), pp. 1–17.
[online] available from http://www.jstor.org/stable/1818907

——— (1962). Reflections on Investment in Man. *The Journal of Political Economy* **70**(5), pp. 1–8.
[online] available from http://www.jstor.org/stable/1829102

——— (1963). *The Economic Value of Education*. New York: John Wiley.

SIANESI, B. and VAN REENEN, J. (2003). The Returns to Education: A Review of the Empirical Macroeconomic Literature. *Journal of Economic Surveys* **17**(2), pp. 157–200.
[online] available from http://dx.doi.org/DOI:10.1111/1467-6419.00192

SMITH, A. (2009). *An Inquiry into the Nature and Causes of the Wealth of Nations*. Digireads.com Publishing.

SPENCE, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics* **87**(3), pp. 355–374.
[online] available from http://www.jstor.org/stable/1882010

STATACORP (2009*a*). *Stata 11 Base Reference Manual*. College Station, TX: Stata Press.

——— (2009*b*). *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.

STATISTICS SOUTH AFRICA (2009). Quarterly Labour Force Survey: Quarter 4, 2008. Statistical Release P0211, Statistics South Africa.

STEEN, L. A. (1990). Numeracy. *Daedalus* **119**(2), pp. 211 – 231.
[online] available from http://www.jstor.org/stable/20025307

UMALUSI (2005). Matric: What is to be Done? Seminar paper, UMALUSI.
[online] available from http://chet.org.za/webfm_send/209

UNESCO (2004). *Education for All (EFA) Global Monitoring Report 2005*. France: UNESCO.
[online] available from http://go.worldbank.org/NOSEUH88U0

VELLA, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources* **33**(1), pp. 127–169.
[online] available from http://www.jstor.org/stable/146317

VON FINTEL, D. and BLACK, P. (2007). From Labour Market Misperceptions to Wage Scarring: The long-term impacts of Youth Unemployment. In *Biennial conference of the Economic Society of South Africa*, Johannesburg, South Africa.

WEBER, E. (2002). An Ambiguous, Contested Terrain: Governance Models for a New South African Education System. *International Journal of Educational Development* **22**, pp. 617–635.
[online] available from http://dx.doi.org/doi:10.1016/S0738-0593(01)00031-1

WEDEGE, T. (2002). Numeracy as a Basic Qualification in Semi-Skilled Jobs. *For the Learning of Mathematics* **22**(3), pp. 23–28.
[online] available from http://www.jstor.org/stable/40248399

WEGNER, L., FLISHER, A. J., CHIKOBVU, P., LOMBARD, C. and KING, G. (2008). Leisure Boredom and High School Dropout in Cape Town, South Africa. *Journal of Adolescence* **31**(3), pp. 421 – 431.
[online] available from http://dx.doi.org/DOI:10.1016/j.adolescence.2007.09.004

WEISS, A. (1995). Human Capital vs. Signalling Explanations of Wages. *The Journal of Economic Perspectives* **9**(4), pp. 133–154.
[online] available from http://www.jstor.org/stable/2138394

WETZELS, C. and ZORLU, A. (2003). Wage Effects of Motherhood: A Double Selection Appriach. NIMA Working Paper No. 22, Nucleo de Investigacao em Microeconomia Aplicada (NIMA), Universidade do Minho.
[online] available from http://econpapers.repec.org/RePEc:nim:nimawp:22

WHITE, H. (1980). Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **48**(4), pp. 817–838.
[online] available from http://www.jstor.org/stable/1912934

WOLPIN, K. I. (1977). Education and Screening. *The American Economic Review* **67**(5), pp. 949–958.
[online] available from http://www.jstor.org/stable/1828076

WOOLDRIDGE, J. (2009). *Introductory Econometrics - A Modern Approach*. Cengage South Western, 4th edn.

WOOLDRIDGE, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 1st edn.

YAMAUCHI, F. (2004). Race, Equity, and Public Schools in Post-Apartheid South Africa: Is Opportunity Eqaul for all Kids? FCND Discussion Paper No. 182, International Food Policy Research Institute (IFPRI), Food Consumption and Nutrition Division (FCND).

——— (2011). School Quality, Clustering and Government Subsidy in Post-apartheid South Africa. *Economics of Education Review* **30**, pp. 146–156.
[online] available from http://dx.doi.org/10.1016/j.econedurev.2010.08.002

Appendix: Regression Tables

Table A.1: Uncorrected Employment Returns to Educational Attainment

	(1)	(2)	(3)	(4)
	Working-Age Sample	*Working-Age Sample*	*Working-Age Sample*	*Working-Age Sample*
Age	0.205***	0.213***	0.149***	0.150***
Age^2	-0.003***	-0.003***	-0.002***	-0.002***
Household Head	0.455***	0.455***	0.446***	0.447***
Female	-0.404***	-0.374***	-0.418***	-0.421***
Married	0.397***	0.304***	0.257***	0.243***
Female × married	-0.316***	-0.275***	-0.235**	-0.224**
Rural Formal	0.147	0.189*	0.282**	0.347***
Tribal Auth Area	-0.415***	-0.400***	-0.313***	-0.255***
Urban Informal	-0.094	-0.118	-0.085	-0.037
Coloured	0.136*	0.087	0.101	0.053
Asian	0.416	0.381	0.237	0.139
White	0.413***	0.336***	0.113	-0.028
Health: Fair		-0.008	-0.028	-0.031
Health: Good		0.139	0.060	0.058
Health: Very Good		0.228***	0.147*	0.148*
Health: Excellent		0.222**	0.107	0.095
Disabled		-0.216***	-0.171**	-0.162**
Emotional Well-being		0.104***	0.097***	0.097***

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)	(4)
Currently Enrolled			-0.723***	-0.825***
Primary			0.080	0.082
Lower Secondary			0.009	0.001
Upper Secondary			0.158*	0.108
Matric			0.414***	0.258**
Diploma/Certificate			0.713***	0.425***
Bachelors Degree			1.045***	0.746***
Postgraduate Degree			0.799***	0.449*
Computer Lit: Basic Use				0.285***
Computer Lit: Highly Literate				0.557***
English competency				0.008
Constant	-3.592***	-3.941***	-2.827***	-2.927***
Observations	15422	13548	13524	13504
P-value	0.000	0.000	0.000	0.000
Area under ROC curve	0.783	0.786	0.803	0.807
Sensitivity	70.801	71.601	73.732	73.592
Specificity	69.627	69.623	69.235	70.058
Cutoff used	0.42	0.42	0.42	0.42
Turning Point - Age	39	40	40	40
Percentile - Age	64	66	64	66

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: probit* command which executes a probit estimation for complex survey data. The sample includes only individuals in the working-age population. Reference categories are as follows: Geographical location (Urban Formal); Race (Black); Health (Poor); Educational Attainment (No Schooling); Computer literacy (Not literate). The chosen cut-off value for the calculated prediction sensitivity and specificity is equal to the proportion of the estimation sample who are employed. Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.2: Uncorrected Employment Returns to Educational Attainment and School Quality

	(1) *Working-Age Sample*	(2) *Quality Sample*	(3) *Quality Sample*	(4) *Quality Sample*
Age	0.150***	0.210***	0.210***	0.209***
Age^2	-0.002***	-0.003***	-0.003***	-0.003***
Household Head	0.447***	0.415***	0.415***	0.393***
Female	-0.421***	-0.396***	-0.396***	-0.393***
Married	0.243***	0.314**	0.313**	0.343**
Female × married	-0.224**	-0.282	-0.281	-0.290*
Rural Formal	0.347***	0.118	0.117	0.171
Tribal Auth Area	-0.255***	-0.209**	-0.208**	-0.247***
Urban Informal	-0.037	-0.102	-0.101	-0.131
Coloured	0.053	0.242**	0.235**	
Asian	0.139	0.348	0.329	
White	-0.028	0.155	0.133	
Health: Fair	-0.031	-0.054	-0.055	-0.052
Health: Good	0.058	0.026	0.026	0.027
Health: Very Good	0.148*	0.160	0.160	0.167
Health: Excellent	0.095	0.041	0.041	0.039
Disabled	-0.162**	0.060	0.060	0.060
Emotional Well-being	0.097***	0.126***	0.127***	0.134***
Currently Enrolled	-0.825***	-0.714***	-0.716***	-0.738***
No Schooling	-0.082			
Lower Secondary	-0.082	-0.074	-0.073	-0.070
Upper Secondary	0.026	0.095	0.096	0.088
Matric	0.176**	0.199	0.200	0.181
Diploma/Certificate	0.343***	0.336	0.337	0.310
Bachelors Degree	0.664***	0.696**	0.696**	0.668*

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)	(4)
Postgraduate Degree	0.367	0.911***	0.910***	0.890***
Computer Lit: Basic Use	0.285***	0.184**	0.183*	0.212**
Computer Lit: Highly Literate	0.557***	0.442***	0.438***	0.470***
English competency	0.008	0.032	0.032	0.035
School Quality			0.137	0.462
Constant	-2.845***	-4.026***	-4.036***	-4.009***
Observations	13504	4573	4573	4575
P-value	0.000	0.000	0.000	0.000
Area under ROC curve	0.807	0.842	0.843	0.841
Sensitivity	73.592	78.280	78.280	78.836
Specificity	70.058	72.515	72.405	72.342
Cutoff used	0.42	0.40	0.40	0.40
Turning Point - Age	40	40	40	40
Percentile - Age	66	82	82	82

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: probit* command which executes a probit estimation for complex survey data. The sample includes only individuals in the working-age population. Reference categories are as follows: Geographical location (Urban Formal); Race (Black); Health (Poor); Educational Attainment (Completed primary); Computer literacy (Not literate). The chosen cut-off value for the calculated prediction sensitivity and specificity is equal to the proportion of the estimation sample who are employed. Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.3: Uncorrected Employment Returns to Educational Attainment and Numeracy

	(1) *Working-Age Sample*	(2) *Numeracy Sample*	(3) *Numeracy Sample*	(4) *Numeracy Sample*
Age	0.150***	0.102***	0.103***	0.118***
Age^2	-0.002***	-0.001***	-0.001***	-0.001***
Household Head	0.447***	0.492***	0.501***	0.512***
Female	-0.421***	-0.544***	-0.540***	-0.516***
Married	0.243***	0.214	0.206	0.223
Female × married	-0.224**	-0.422**	-0.416**	-0.430**
Rural Formal	0.347***	0.383***	0.385***	0.350**
Tribal Auth Area	-0.255***	-0.062	-0.076	-0.085
Urban Informal	-0.037	-0.008	0.012	-0.002
Coloured	0.053	0.249**	0.230**	0.219**
Asian	0.139	1.766***	1.754***	1.685***
White	-0.028	0.270	0.237	0.211
Health: Fair	-0.031	0.039	-0.020	0.002
Health: Good	0.058	0.110	0.046	0.066
Health: Very Good	0.148*	0.117	0.056	0.074
Health: Excellent	0.095	0.029	-0.031	0.004
Disabled	-0.162**	-0.004	-0.014	-0.063
Emotional Well-being	0.097***	0.085*	0.086*	0.087**
Currently Enrolled	-0.825***	-0.745***	-0.756***	-0.797***
Primary	0.082	0.313	0.243	
Lower Secondary	0.001	0.337	0.254	
Upper Secondary	0.108	0.431	0.328	
Matric	0.258**	0.615	0.509	
Diploma/Certificate	0.425***	0.780	0.671	
Bachelors Degree	0.746***	1.210	1.091	

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)	(4)
Postgraduate Degree	0.449*	0.529	0.379	
Computer Lit: Basic Use	0.285***	0.224**	0.228**	0.315***
Computer Lit: Highly Literate	0.557***	0.268*	0.264*	0.403***
English competency	0.008	-0.002	-0.004	0.025
Numeracy			0.089**	0.097***
Constant	-2.927***	-2.662***	-2.447***	-2.473***
Observations	13504	3395	3395	3404
P-value	0.000	0.000	0.000	0.000
Area under ROC curve	0.807	0.840	0.841	0.837
Sensitivity	73.592	78.504	78.762	78.938
Specificity	70.058	73.432	73.387	73.614
Cutoff used	0.42	0.34	0.34	0.34
Turning Point - Age	40	52	52	48
Percentile - Age	66	97	97	95

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: probit* command which executes a probit estimation for complex survey data. The sample includes only individuals in the working-age population. Reference categories are as follows: Geographical location (Urban Formal); Race (Black); Health (Poor); Educational Attainment (None); Computer literacy (Not literate). The chosen cut-off value for the calculated prediction sensitivity and specificity is equal to the proportion of the estimation sample who are employed. Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.4: Uncorrected Earnings returns to Educational Attainment

	(1)	(2)	(3)	(4)	(5)
	Earnings Sample	*Earnings Sample*	*Earnings Sample*	*Earnings Sample*	*Earnings Sample*
Age	0.062***	0.059***	0.061***	0.066***	0.064***
Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
Household Head	0.215***	0.176***	0.179***	0.182***	0.166***
Female	-0.441***	-0.328***	-0.318***	-0.338***	-0.327***
Married	0.343***	0.249***	0.240***	0.228***	0.204***
Rural	-0.230***	-0.235***	-0.231***	-0.166***	-0.148***
Coloured	0.170**	0.112*	0.099*	0.062	0.082
Asian	0.593**	0.607***	0.608***	0.483**	0.494**
White	0.735***	0.612***	0.595***	0.465***	0.532***
Primary	0.118*	0.117*	0.121*	0.038	0.016
Lower Secondary	0.401***	0.296***	0.298***	0.144*	0.109
Upper Secondary	0.611***	0.511***	0.508***	0.269***	0.230**
Matric	1.073***	0.837***	0.828***	0.456***	0.421***
Diploma/Certificate	1.607***	1.244***	1.235***	0.809***	0.697***
Bachelors Degree	2.040***	1.460***	1.459***	1.031***	0.878***
Postgraduate Degree	2.230***	1.774***	1.763***	1.296***	1.198***
Semi-skilled		0.264***	0.263***	0.216***	0.177***
Skilled		0.509***	0.508***	0.408***	0.398***
Self-Employed		-0.690***	-0.691***	-0.714***	-0.655***
× Semi-skilled		0.595**	0.604**	0.636***	0.672***
× Skilled		0.599**	0.615**	0.678***	0.744***
Casually Employed		-0.691***	-0.670***	-0.662***	-0.608***
× Semi-skilled		-0.030	-0.034	0.001	0.035
× Skilled		0.337**	0.316**	0.248*	0.208*
Emotional Well-being			0.056***	0.052***	0.055***

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)	(4)	(5)
Computer Lit: Basic Use				0.260***	0.249***
Computer Lit: Highly Literate				0.443***	0.445***
English competency				0.050***	0.044***
Union member					0.354***
Constant	5.281***	5.407***	5.370***	5.213***	5.284***
Observations	5728	4552	4543	4537	4452
R-squared	0.516	0.581	0.583	0.599	0.617
F Statistic	145.022	133.961	131.620	140.604	141.812
Turning Point - Age	47	48	48	48	46
Percentile - Age	75	78	78	78	76

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: regress* command which executes OLS estimation for complex survey data. The dependent variable is the *log of monthly earnings* and the sample includes only individuals in the working-age population. Reference categories are as follows: Race (Black); Educational Attainment (No Schooling); Occupational skill level (Unskilled); Computer literacy (Not literate). Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.5: Uncorrected Earnings returns to Educational Attainment and School Quality

	(1) *Earnings Sample*	(2) *Quality Sample*	(3) *Quality Sample*	(4) *Quality Sample*
Age	0.064***	0.072***	0.073***	0.065***
Age^2	-0.001***	-0.001***	-0.001***	-0.001***
Household Head	0.166***	0.214***	0.214***	0.199***
Female	-0.327***	-0.294***	-0.295***	-0.312***
Married	0.204***	0.229***	0.228***	0.257***
Rural	-0.148***	-0.135**	-0.137**	-0.133**
Coloured	0.082	0.074	0.057	
Asian	0.494**	0.778**	0.728**	
White	0.532***	0.582***	0.525***	
Semi-skilled	0.177***	0.249***	0.250***	0.255***
Skilled	0.398***	0.498***	0.499***	0.556***
Self-Employed	-0.655***	-0.722***	-0.725***	-0.783***
× Semi-skilled	0.672***	0.629***	0.632***	0.664***
× Skilled	0.744***	0.214	0.226	0.344
Casually Employed	-0.608***	-0.406***	-0.406***	-0.430***
× Semi-skilled	0.035	-0.028	-0.030	0.021
× Skilled	0.208*	-0.126	-0.115	-0.059
Union member	0.354***	0.337***	0.337***	0.294***
Emotional Well-being	0.055***	0.048*	0.050*	0.061**
Computer Lit: Basic Use	0.249***	0.347***	0.343***	0.383***
Computer Lit: Highly Literate	0.445***	0.496***	0.487***	0.570***
English competency	0.044***	0.032	0.031	0.040*
No Schooling	-0.016			
Lower Secondary	0.093	-0.015	-0.014	-0.060
Upper Secondary	0.214***	0.054	0.058	0.036

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)	(4)
Matric	0.405***	0.217	0.222	0.187
Diploma/Certificate	0.681***	0.545***	0.552***	0.494***
Bachelors Degree	0.862***	0.563***	0.569***	0.519***
Postgraduate Degree	1.182***	0.874***	0.874***	0.847***
School Quality			0.339	1.599***
Constant	5.300***	5.229***	5.203***	5.270***
Observations	4452	1497	1497	1498
R-squared	0.617	0.611	0.611	0.593
F Statistic	141.812	94.291	90.994	69.870
Turning Point - Age	46	45	45	47
Percentile - Age	76	84	84	87

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: regress* command which executes OLS estimation for complex survey data. The dependent variable is the *log of monthly earnings* and the sample includes only individuals in the working-age population. Reference categories are as follows: Race (Black); Educational Attainment (Primary Education); Occupational skill level (Unskilled); Computer literacy (Not literate). Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.6: Uncorrected Earnings returns to Educational Attainment and Numeracy

	(1) *Earnings Sample*	(2) *Numeracy Sample*	(3) *Numeracy Sample*	(4) *Numeracy Sample*
Age	0.064***	0.045	0.046	0.065**
Age^2	-0.001***	-0.000	-0.000	-0.001*
Household Head	0.166***	0.068	0.070	0.070
Female	-0.327***	-0.516***	-0.517***	-0.480***
Married	0.204***	0.253***	0.246***	0.243***
Rural	-0.148***	-0.107	-0.110	-0.095
Coloured	0.082	0.023	0.010	-0.042
Asian	0.494**	0.271	0.275	0.342
White	0.532***	0.586***	0.559***	0.628***
Semi-skilled	0.177***	0.051	0.047	0.100
Skilled	0.398***	0.187*	0.187*	0.438***
Self-Employed	-0.655***	-0.936***	-0.929***	-0.958***
× Semi-skilled	0.672***	0.781**	0.780**	0.825**
× Skilled	0.744***	1.491***	1.478***	1.577***
Casually Employed	-0.608***	-0.894***	-0.885***	-0.849***
× Semi-skilled	0.035	0.161	0.157	0.160
× Skilled	0.208*	0.682***	0.632***	0.403**
Union member	0.354***	0.279***	0.271***	0.369***
Emotional Well-being	0.055***	-0.001	0.001	0.028
Computer Lit: Basic Use	0.249***	0.106	0.114	0.276***
Computer Lit: Highly Literate	0.445***	0.395***	0.395***	0.680***
English competency	0.044***	0.059*	0.059*	0.098***
Primary	0.016	0.265**	0.207*	
Lower Secondary	0.109	0.097	0.027	
Upper Secondary	0.230**	0.335**	0.256	

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)	(4)
Matric	0.421***	0.630***	0.551***	
Diploma/Certificate	0.697***	0.985***	0.896***	
Bachelors Degree	0.878***	1.363***	1.258***	
Postgraduate Degree	1.198***	1.329***	1.222***	
Numeracy			0.048	0.081**
Constant	5.284***	5.578***	5.666***	5.441***
Observations	4452	885	885	890
R-squared	0.617	0.653	0.654	0.617
F Statistic	141.812	80.730	97.802	43.037
Turning Point - Age	46	66	66	52
Percentile - Age	76	100	100	95

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: regress* command which executes OLS estimation for complex survey data. The dependent variable is the *log of monthly earnings* and the sample includes only individuals in the working-age population. Reference categories are as follows: Race (Black); Educational Attainment (No Schooling); Occupational skill level (Unskilled); Computer literacy (Not literate). Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.7: Baseline Selection Equations for the Participant, School Quality and Numeracy Samples

	(1)	(2)	(3)
	Labour Force Participation	*School Quality Match*	*Numeracy Test Participation*
Age	0.170***	-0.012***	-0.052***
Age^2	-0.002***		0.000***
Household Head	0.092*		
Household Size	-0.024		-0.017**
Household $Size^2$	0.003**		
Female	-0.295***	0.124***	
Married	0.301***		
$\times$ Female	-0.581***		
Rural Formal	0.040		-0.140
Tribal Auth Area	-0.317***		-0.236**
Urban Informal	0.219*		0.171
Coloured		0.628***	-0.004
Asian		0.563*	-0.301
White		0.331	-0.297*
Grant Recipient	-0.262***		
Health: Fair	0.212**		0.308**
Health: Good	0.360***		0.333**
Health: Very Good	0.332***		0.259**
Health: Excellent	0.356***		0.299**
Disability	-0.350***		
Primary	0.145**		1.128***
Lower Secondary	0.248***		1.299***
Upper Secondary	0.314***		1.398***
Matric	0.613***		1.489***

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)
Diploma/Certificate	0.915***		1.558***
Bachelors Degree	0.903***		1.662***
Postgraduate Degree	0.495*		2.004***
Currently Enrolled	-1.567***	0.148***	
Education		0.485***	
Education^2		-0.017***	
Isindebele		0.602**	
Isixhosa		0.334	
Isizulu		0.500**	
Sepedi		0.602***	
Sesotho		0.607***	
Setswana		0.559**	
Siswati		0.609**	
Tshivenda		0.816***	
Xitsonga		0.304	
Afrikaans		-0.074	
Father Education Info		0.187***	
Mother Education Info		0.085**	
HH Quality Response Rate		1.500***	
HH Quality Response Rate^2		-0.949***	
Emotional Well-being			0.030
Emotional Well-being^2			0.038*
Time before Test			-0.005***
Time before Test^2			0.000**
Took measurements			0.975***
HH Test Response Rate			1.218***

(continued on next page)

(continued from previous page)

	(1)	(2)	(3)
Constant	-2.125***	-4.013***	-1.844***
Observations	14948	15288	13440
P-value	0.000	0.000	0.000
Area under ROC curve	0.859	0.800	0.835
Sensitivity	83.830	83.999	79.160
Specificity	71.087	61.327	70.382
Cutoff used	0.62	0.31	0.25
Turning Point - Age	37		84
Percentile - Age	59		100
Turning Point - Household Size	4		
Percentile - Household Size	50		
Turning Point - Educ		14	
Percentile - Educ		98	
Turning Point - Household Quality Response Rate		1	
Percentile - Household Quality Response Rate		88	
Turning Point - Emotion Index			-0
Percentile - Emotion Index			33
Turning Point - Time Before Test			339
Percentile - Time Before Test			100

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: probit* command which executes a probit estimation for complex survey data. The dependent variables for the respective probit selection models are labour force participation, having school quality data available, and participation in the numeracy test module. Reference categories are as follows: Race (Black); Educational Attainment (No Schooling); Language (English); Geographical location (Urban Formal). Turning points are calculated based on the coefficient estimates: $-\beta_x/2\beta_{x^2}$ and the percentile statistics report the percentiles corresponding to the turning point in question.

Table A.8: Corrected Employment Returns to Educational Attainment and School Quality

	(i)*	(ii)*	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Age	0.225***	0.062***	0.071***	0.156***	0.156***	0.152***	0.278***	0.278***	0.281***
Age^2	-0.003***	-0.000**	-0.000**	-0.002***	-0.002***	-0.002***	-0.003***	-0.003***	-0.003***
Female	-0.604***	-0.518***	-0.534***	-0.530***	-0.530***	-0.531***	-0.539***	-0.539***	-0.539***
Rural Formal	0.346***	0.412***	0.416***	0.133	0.131	0.156	0.138	0.139	0.212*
Tribal Auth Area	-0.330***	-0.223***	-0.236***	-0.208*	-0.207*	-0.252**	-0.277***	-0.277***	-0.314***
Urban Informal	-0.060	-0.185**	-0.179**	-0.251**	-0.249**	-0.288***	-0.112	-0.112	-0.142
Coloured	0.027	0.036	0.036	0.239*	0.224*		0.226**	0.229**	
Asian	0.115	0.196	0.197	0.356	0.322		0.375	0.383	
White	-0.077	0.053	0.049	0.418	0.376		0.063	0.073	
Emotional Well-being	0.094***	0.115***	0.115***	0.129***	0.130***	0.141***	0.111***	0.111***	0.118***
No Schooling	-0.163**	-0.007	-0.017						
Lower Secondary	-0.069	-0.132*	-0.128*	-0.042	-0.039	-0.022	-0.002	-0.003	-0.026
Upper Secondary	0.057	0.033	0.041	0.236	0.240	0.255	0.239	0.237	0.196
Matric	0.320***	0.136	0.155*	0.293	0.298	0.312	0.455*	0.453*	0.392
Diploma/Certificate	0.523***	0.268**	0.296***	0.380	0.386	0.396	0.615**	0.613**	0.546**
Bachelors Degree	0.757***	0.615**	0.642**	0.626	0.634	0.693	0.938**	0.936**	0.846**
Postgraduate Degree	0.559**	1.504***	1.520***	1.774***	1.778***	1.816***	1.256***	1.256***	1.189***
Computer Lit: Basic Use	0.191***	0.170**	0.167**	0.164	0.161	0.192*	0.103	0.103	0.128

(continued on next page)

(continued from previous page)

	(i)*	(ii)*	(1)^a	(2)^a	(3)^a	(4)^a	(5)^b	(6)^b	(7)^b
Computer Lit: Highly Literate	0.434***	0.512***	0.505***	0.345**	0.336**	0.386***	0.291**	0.293**	0.316**
English competency	-0.007	-0.014	-0.014	-0.001	-0.002	-0.001	0.009	0.009	0.014
School Quality					0.258	0.876*		-0.058	0.234
Constant	-4.017***	-0.898***	-1.082***	-2.810***	-2.823***	-2.786***	-5.276***	-5.269***	-5.263***
atanh $\rho_{Participation}$			0.078	0.113	0.111	0.113			
atanh $\rho_{Quality}$							0.101	0.099	0.065
N: Employment	13863	8298	8298	2873	2873	2873	4623	4623	4623
N: Participation			13908	8483	8483	8483			
N: Quality							15269	15269	15269
P-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\rho_{Participation}$			0.0776	0.1127	0.1105	0.1128			
$\rho_{School\ Quality}$							0.1004	0.0989	0.0649

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: heckprob* command which executes the ML binary second-stage outcome version of the Heckman (1979) estimation procedure for complex survey data. Reference categories are as follows: Race (Black); Educational Attainment (Completed Primary); Computer Literacy (None); Geographical location (Urban Formal). The number of observations reported for each equation correspond to the number of uncensored observations available for the estimation sample in question. Models (i) and (ii) do not control for any sample selection but respectively show the uncorrected outcomes for the full sample and the censored participant sample in column (1). Models (2), (3) and (4) use the school quality sample, corrected for selection into the participant sample and models (5), (6) and (7) use the school quality sample when correcting for selection into the school quality sample.

* Does not control for any sample selection

^{a} Controls for selection into participant sample (LFP)

^{b} Controls for selection into school quality

Table A.9: Corrected Employment Returns to Educational Attainment and Numeracy

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Age	0.071***	0.079**	0.079*	0.067*	0.205***	0.207***	0.234***
Age^2	-0.000**	-0.000	-0.000	-0.000	-0.002***	-0.002***	-0.003***
Female	-0.534***	-0.791***	-0.783***	-0.749***	-0.713***	-0.707***	-0.691***
Rural Formal	0.416***	0.409**	0.411**	0.420**	0.291**	0.294**	0.264*
Tribal Auth Area	-0.236***	-0.095	-0.106	-0.062	-0.213*	-0.226**	-0.241**
Urban Informal	-0.179**	-0.181*	-0.160	-0.190*	0.018	0.038	0.009
Coloured	0.036	0.048	0.034	0.026	0.208*	0.194*	0.152
Asian	0.197	1.820***	1.809***	1.776***	2.384***	2.358***	2.264***
White	0.049	0.105	0.071	0.115	0.147	0.115	0.115
Emotional Well-being	0.115***	0.107**	0.107**	0.108**	0.080*	0.079*	0.089**
Primary	0.017	1.178*	1.072		0.573	0.503	
Lower Secondary	-0.111	1.132*	1.010		0.661	0.585	
Upper Secondary	0.057	1.315*	1.171		0.728	0.634	
Matric	0.172	1.396*	1.254*		1.023	0.928	
Diploma/Certificate	0.313*	1.477**	1.325*		1.196*	1.093	
Bachelors Degree	0.659**	7.505***	7.684***		1.618**	1.500*	
Postgraduate Degree	1.537***	2.698***	2.520***		0.804	0.674	
Computer Lit: Basic Use	0.167**	0.145	0.152	0.214*	0.112	0.115	0.227**

(continued on next page)

(continued from previous page)

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Computer Lit: Highly Literate	0.505***	0.382**	0.377**	0.519***	0.115	0.109	0.270*
English competency	-0.014	-0.018	-0.019	0.006	-0.013	-0.015	0.025
Numeracy			0.081	0.099**		0.075*	0.087**
Constant	-1.098***	-2.568**	-2.388**	-1.080*	-4.680***	-4.567***	-4.456***
atanh $\rho_{Participation}$	0.078	0.129	0.127	-0.018			
atanh $\rho_{Numeracy}$					0.199*	0.210*	0.156
N: Earnings	8298	1921	1921	1921	3278	3278	3278
N: Employment	13908	7531	7531	7531			
N: Numeracy					13407	13407	13407
P-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\rho_{Participation}$	0.0776	0.1283	0.1263	-0.0181			
$\rho_{Numeracy}$					0.1965	0.2072	0.1549

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: heckprob* command which executes the ML binary second-stage outcome version of the Heckman (1979) estimation procedure for complex survey data. Reference categories are as follows: Race (Black); Educational Attainment (No Schooling); Computer Literacy (None); Geographical location (Urban Formal). The number of observations reported for each equation correspond to the number of uncensored observations available for the estimation sample in question. Model (1) uses the participant sample. Models (2), (3) and (4) use the numeracy sample, corrected for selection into the participant sample and models (5), (6) and (7) use the numeracy sample when correcting for selection into the numeracy sample.

^{a} Controls for selection into participant sample (LFP)

^{b} Controls for selection into numeracy test participation

Table A.10: Corrected Earnings Returns to Educational Attainment and School Quality

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Age	0.050***	0.008	0.007	-0.026	0.076***	0.077***	0.071***
Age^2	-0.001***	0.000	0.000	0.001	-0.001***	-0.001***	-0.001***
Household Head	0.126***	0.060	0.056	-0.008	0.216***	0.217***	0.202***
Female	-0.284***	-0.169	-0.166	-0.128*	-0.294***	-0.295***	-0.319***
Married	0.195***	0.208***	0.206***	0.212***	0.222***	0.221***	0.254***
Rural	-0.130***	-0.082	-0.082	-0.046	-0.135**	-0.137**	-0.130**
Coloured	0.066	-0.025	-0.046		0.063	0.044	
Asian	0.479**	0.616	0.554		0.771**	0.715**	
White	0.512***	0.529***	0.461***		0.611***	0.547***	
Semi-Skilled	0.186***	0.239***	0.240***	0.236***	0.260***	0.261***	0.266***
Skilled	0.414***	0.525***	0.526***	0.567***	0.515***	0.516***	0.576***
Self-Employed	-0.664***	-0.671**	-0.673***	-0.655***	-0.719***	-0.724***	-0.784***
× Semi-skilled	0.673***	0.594***	0.598***	0.618***	0.620***	0.624***	0.656***
× Semi-skilled	0.750***	0.171	0.183	0.240	0.192	0.195	0.255
Casually Employed	-0.584***	-0.349*	-0.347*	-0.341**	-0.396***	-0.396***	-0.412***
× Semi-skilled	0.022	-0.028	-0.028	0.032	-0.039	-0.041	0.004
× Semi-skilled	0.220*	-0.024	-0.008	0.083	-0.152	-0.139	-0.123
Union member	0.346***	0.331***	0.332***	0.299***	0.328***	0.329***	0.285***

(continued on next page)

(continued from previous page)

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Emotional Well-being	0.047**	0.016	0.017	0.014	0.050*	0.052*	0.062**
Computer Lit: Basic Use	0.235***	0.301***	0.295***	0.301***	0.357***	0.353***	0.389***
Computer Lit: Highly Literate	0.419***	0.374*	0.361**	0.395***	0.509***	0.500***	0.573***
English competency	0.043***	0.015	0.013	0.013	0.033	0.032	0.041*
No Schooling	-0.012						
Lower Secondary	0.091	-0.171	-0.174	-0.248*	-0.034	-0.035	-0.020
Upper Secondary	0.215***	-0.242	-0.245	-0.359**	0.033	0.036	0.111
Matric	0.393***	-0.122	-0.125	-0.258	0.194	0.198	0.280
Diploma/Certificate	0.665***	0.158	0.155	-0.016	0.510**	0.514**	0.578**
Bachelors Degree	0.816***	0.135	0.129	-0.030	0.509*	0.512*	0.609*
Postgraduate Degree	1.164***	0.473	0.463	0.318	0.821***	0.818***	0.932**
School Quality			0.390	1.277***		0.384	1.655***
Constant	5.650***	7.260**	7.287**	8.126***	5.188***	5.158***	4.958***
atanh $\rho_{Employment}$	-0.186	-0.642	-0.661	-0.910***			
atanh $\rho_{Quality}$					-0.021	-0.024	0.106
$\ln \sigma_{Employment}$	-0.296***	-0.220	-0.215	-0.122			
$\ln \sigma_{Quality}$					-0.326***	-0.327***	-0.300***

(continued on next page)

(continued from previous page)

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
N: Earnings	4362	1472	1472	1472	1478	1478	1478
N: Employment	12254	9364	9364	9364			
N: School Quality					12124	12124	12124
P-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\rho_{Employment}$	-0.1834	-0.5659	-0.5791	-0.7209			
$\rho_{School\ Quality}$					-0.0210	-0.0236	0.1058
$\sigma_{Employment}$	0.7438	0.8021	0.8062	0.8849			
$\sigma_{School\ Quality}$					0.7219	0.7214	0.7409
$\lambda_{Employment}$	-0.1364	-0.4540	-0.4669	-0.6380			
$\lambda_{School\ Quality}$					-0.0152	-0.0170	0.0784

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: heckman* command which executes the ML version of the Heckman (1979) estimation procedure for complex survey data. The dependent variable in the outcome equation is the *log of monthly earnings*. Reference categories are as follows: Race (Black); Educational Attainment (Completed Primary); Occupational skill level (Unskilled); Computer Literacy (None); Geographical location (Urban Formal). The number of observations reported for each equation correspond to the number of uncensored observations available for the estimation sample in question. Model (1) uses the full earnings sample. Models (2), (3) and (4) use the school quality sample, corrected for selection into the earnings sample and models (5), (6) and (7) use the school quality sample when correcting for selection into the school quality sample.

[a] Controls for selection into earnings sample (employment)

[b] Controls for selection into school quality sample

Table A.11: Corrected Earnings Returns to Educational Attainment and Numeracy

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Age	0.050***	0.014	0.016	-0.030	0.042	0.043	0.045
Age^2	-0.001***	0.000	0.000	0.001	-0.000	-0.000	-0.000
Household Head	0.126***	-0.030	-0.024	-0.229**	0.051	0.055	0.050
Female	-0.284***	-0.425***	-0.430***	-0.220*	-0.530***	-0.530***	-0.512***
Married	0.195***	0.235***	0.230***	0.194**	0.245***	0.240***	0.214**
Rural	-0.130***	-0.059	-0.064	0.073	-0.073	-0.077	0.003
Coloured	0.066	-0.003	-0.014	-0.122	0.003	-0.008	-0.059
Asian	0.479**	0.296	0.297	0.381	0.315	0.315	0.409
White	0.512***	0.608***	0.583***	0.686***	0.639***	0.611***	0.772***
Semi-Skilled	0.186***	0.084	0.079	0.091	0.038	0.036	0.072
Skilled	0.414***	0.232**	0.231**	0.396***	0.182*	0.182*	0.388***
Self-Employed	-0.664***	-0.883***	-0.877***	-0.765***	-0.979***	-0.969***	-0.952***
× Semi-skilled	0.673***	0.753**	0.752**	0.790**	0.769**	0.762**	0.761**
× Semi-skilled	0.750***	1.419***	1.409***	1.349***	1.433***	1.426***	1.442***
Casually Employed	-0.584***	-0.849***	-0.843***	-0.758***	-0.859***	-0.853***	-0.818***
× Semi-skilled	0.022	0.141	0.138	0.180	0.113	0.107	0.094
× Semi-skilled	0.220*	0.589**	0.548**	0.245	0.633**	0.610**	0.411*
Union member	0.346***	0.267***	0.260***	0.344***	0.280***	0.269***	0.362***

(continued on next page)

(continued from previous page)

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
Emotional Well-being	0.047**	-0.024	-0.021	-0.038	0.005	0.007	0.021
Computer Lit: Basic Use	0.235***	0.081	0.089	0.136	0.135	0.145	0.277***
Computer Lit: Highly Literate	0.419***	0.349**	0.351**	0.459***	0.440***	0.438***	0.683***
English competency	0.043***	0.056	0.056	0.058**	0.057	0.056	0.082***
Primary	0.012	0.019	-0.023		0.136	0.080	
Lower Secondary	0.103	-0.198	-0.249		-0.071	-0.136	
Upper Secondary	0.226**	0.053	-0.006		0.180	0.107	
Matric	0.405***	0.306	0.249		0.417	0.346	
Diploma/Certificate	0.676***	0.661*	0.594*		0.777***	0.694**	
Bachelors Degree	0.828***	0.992**	0.913**		1.189***	1.075***	
Postgraduate Degree	1.176***	0.932**	0.853**		1.069***	0.974***	
Numeracy			0.043	0.052		0.048	0.077*
Constant	5.638***	6.663***	6.696***	8.111***	5.983***	6.057***	6.248***
atanh $\rho_{Employment}$	-0.186	-0.294	-0.282	-0.940***			
atanh $\rho_{Numeracy}$					-0.143	-0.138	-0.363*
$\ln \sigma_{Employment}$	-0.296***	-0.301***	-0.304***	-0.079			
$\ln \sigma_{Numeracy}$					-0.315***	-0.317***	-0.239***

(continued on next page)

(continued from previous page)

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b]	(6)[b]	(7)[b]
N: Earnings	4362	876	876	876	838	838	838
N: Employment	12254	8768	8768	8768			
N: Numeracy					10967	10967	10967
P-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\rho_{Employment}$	-0.1834	-0.2862	-0.2752	-0.7351			
$\rho_{Numeracy}$					-0.1419	-0.1375	-0.3476
$\sigma_{Employment}$	0.7438	0.7401	0.7375	0.9237			
$\sigma_{Numeracy}$					0.7298	0.7283	0.7873
$\lambda_{Employment}$	-0.1364	-0.2118	-0.2029	-0.6790			
$\lambda_{Numeracy}$					-0.1036	-0.1001	-0.2736

NOTES: *Significant at the 10% level **Significant at the 5% level *** Significant at the 1% level. Significance levels are based on linearised robust standard errors. All regressions are estimated using Stata/SE 11.2's *svy: heckman* command which executes the ML version of the Heckman (1979) estimation procedure for complex survey data. The dependent variable in the outcome equation is the *log of monthly earnings*. Reference categories are as follows: Race (Black); Educational Attainment (No Schooling); Occupational skill level (Unskilled); Computer Literacy (None); Geographical location (Urban Formal). The number of observations reported for each equation correspond to the number of uncensored observations available for the estimation sample in question. Model (1) uses the full earnings sample. Models (2), (3) and (4) use the numeracy sample, corrected for selection into the earnings sample and models (5), (6) and (7) use the numeracy sample when correcting for selection into the numeracy sample.

[a] Controls for selection into earnings sample (employment)

[b] Controls for selection into numeracy test participation