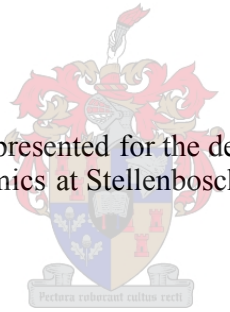
The open door of learning – Access restricted: School effectiveness and efficiency across the South African education system

by
Debra Lynne Shepherd



Dissertation presented for the degree of Doctor of Economics at Stellenbosch University

Promoters: Prof Chris Elbers
Faculty of Economics and Business Administration
(Vrije University Amsterdam)

Prof Servaas Van der Berg
Faculty of Economics and Management Science
(Stellenbosch University)

December 2016

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 2016

"The doors of learning and culture shall be opened!"

— *The Freedom Charter, 1955*

"If your plan is for one year, plant rice. If your plan is for ten years, plant trees. If your plan is for one hundred years, educate children. "

— *Confucius*

Appreciation and Acknowledgements

The completion of this thesis would not have been possible without the kind financial support of the National Research Foundation and Vrije University as co-contributors to the VU-NRF Desmond Tutu Doctoral Scholarship programme. I would also like to thank the Economics Department of Stellenbosch University (in particular Andrie Schoombee), the Postgraduate Office of Stellenbosch University, the Department of Development Economics at Vrije University (in particular Trudi Heemskerk) and SAVUSA for their consistent support.

This has been a long journey. Sometimes it is hard to believe that I have spent as many years (13 to be exact) in higher education as I spent in primary and high school. I doubt that I would have been willing to commit myself to so much education if it weren't for the inspiring and thought-provoking teachers, friends and colleagues that I have had the pleasure of coming into contact with along the way. Economics was something I fell into by accident when I decided that Engineering was not my calling (sorry, Dad). My initial love for the subject was sparked by my A-levels teacher, Paul Pemberton, and this appreciation was only further nurtured by my undergraduate and postgraduate lecturers, in particular Sarel Steele, Philip Black, Rulof Burger, Stan du Plessis, Albert van der Merwe, Sampie Terreblanche and Estian Calitz.

I would also like to thank a number of people whom have offered invaluable feedback and comments on parts of this dissertation: Nic Spaull, Ronelle Burger, Paula Armstrong, Hendrik van Broekhuizen, Gabrielle Wills, Martin Gustafsson, Stephen Taylor, Laura Rossouw, Cobus Burger, Dieter von Fintel, Marisa von Fintel, Anja Smith, Omphile Ramela, Neil Rankin, Volker Schoer, Stefan Klaasen, Nwabisa Makaluza, Murray Leibbrandt, Lucasz Marc, Lizette Swart, Berber Kramer, Wendy Janssens, Menno Pradhan, Maarten Lindeboom, Melinda Vigh, Zlata Tanović and Fujin Zhou. A special thank you to my fourth floor colleagues, Eldridge Moses and Kholekile Malindi; I am so grateful to you for allowing me the space to vent and laugh when I so desperately needed to over the past couple of years.

Chapters from this thesis have been presented at departments and conferences both nationally and internationally, and I would like to thank commentators from the following institutions and groups for their feedback: Stellenbosch University Department of Economics, Research in Social Economic Policy (ReSEP), the African Microeconomic Research Unit (AMERU), University of Cape Town, University of the Witwatersrand, North-West University, University of Pretoria, HSRC, University of Leuven (LEER Institute), Bath University, Vrije University, Tinbergen Institute, attendees of the IEA

International Research Conference (2015) and attendees of the International Workshop on Applied Economics of Education (2014).

This thesis would not have been possible without the support of my two supervisors, Chris Elbers and Servaas van der Berg. To Chris: thank you for welcoming me into your department at Vrije University. I am forever grateful for the opportunity to acquire knowledge through the TI and the VU. I can truly say that I have felt nothing but at home in the Development Economics department and in Amsterdam. I will use whatever excuse necessary to keep visiting and look forward to all our future collaborations together. To Servaas: your guidance over the last decade has been nothing less than extraordinary. You have been so patient and giving of your time and expertise. Thank you for giving me every opportunity to be challenged and ultimately shine in my work.

I am in a fortunate position to be surrounded and supported by my wonderful family and closest friends. Thank you to my parents, Brian and Meg Shepherd, for the love, time and effort invested in me. To my two beautiful and inspirational sisters, Jessica and Jene', I am in awe of your work ethic and grace. I am blessed to be surrounded by strong and caring women (I'm looking at you Diana, Emma and Olivia).

Last, but definitely not least, to Lee.

Table of Contents

1 Introduction 13

1.1 Background 13
1.2 Dominant Approaches to Understanding Education Quality 15
1.2.1 Human capital approach 15
1.2.2 Human rights approach 16
1.2.3 Criticisms of the human capital and human rights approaches 17
1.3 Education Policy in South Africa Post-Apartheid 18
1.3.1 Focus on Equity, Redress and Access 19
1.3.2 Teacher Interventions and Curriculum Reform 23
1.3.3 Policy Lessons 24
1.4 Social Justice Approach to Education Quality 26
1.5 Thesis Outline 29

2 Tree of knowledge: A nonparametric approach to modelling primary school outcomes in South Africa 37

2.1 Introduction 37
2.2 The Performance of Disadvantaged Schools in South Africa 41
2.3 Empirical approach 43
2.3.1 Regression trees 43
2.3.2 Boosting and regularisation 44
2.3.3 Interpretation 46
2.4. Data 47
2.5 Results 48
2.5.1 From single to multiple tree regressions 48
2.5.2 Tuning of model parameters 49
2.5.3 Relative influence of control variables and partial dependence plots 51
2.5 Identification of important interactions 56
2.6 Robustness checks 61
2.6.1 Sensitivity to dropping observations 61
2.6.2 Comparisons across independent data sets 62
2.6.3 Comparisons with alternative modelling approaches 63
2.6.4 Regression Tree Modelling with Clustered Data 66
2.7 Concluding Remarks 69
Appendix to Chapter 2 71

3 A question of efficiency: decomposing South African reading test scores using PIRLS 2006 75

3.1 Introduction 75
3.2 Methodology 77
3.2.1 Oaxaca-Blinder decomposition 77
3.2.2 Semi-parametric decomposition 80
3.3 Data and summary statistics 85
3.4 Empirical results 89

3.4.1 Aggregate decomposition results 89
3.4.2 Sensitivity checks 92
3.5 Concluding remarks 99
Appendix to Chapter 3 101

4 Balancing Act: A Semi-parametric approach for determining the local treatment effect of school type with an application to South Africa 107
4.1 Introduction 107
4.2 Motivation and Methodological Approach 110
4.2.1 Study Design: The effect of attending a former advantaged school 110
4.2.2 Potential Outcomes Framework 112
4.2.3 Inducing randomness: Creating comparable school and student groups 113
4.2.4 Post-matching estimation strategy 119
4.3 Data description 120
4.4 Empirical results 123
4.4.1 Matching students 123
4.4.2 Matching schools 125
4.4.3 Estimates of the treatment effect 128
4.4.4 Sensitivity analysis 131
4.5 Regression meets matching and propensity score weighting 132
5. Concluding remarks 136
Appendix for chapter 4 138

5 Learn to teach, teach to learn: A within-pupil across-subject approach to estimating the impact of teacher subject knowledge on South African grade 6 performance 139
5.1 Introduction 139
5.2 Policy context and previous findings in South Africa 141
5.3 Data and Descriptive Statistics 145
5.4 Estimation strategy: correlated random errors model 147
5.5 Results 151
5.5.1 Base results 151
5.5.2 Heterogeneous effects across student sub-groups 152
5.5.3 Returns to teacher and classroom characteristics 161
5.5.4 Correction for teacher unobservables 162
5.5.5 Fixed effects estimation 165
5.6 Conclusion 166
Appendix to Chapter 5 170

6 Compulsory tutorial programmes and performance in undergraduate microeconomics: A regression discontinuity design 175
6.1 Introduction 175
6.2 Overview of the literature 176

6.3 Data and Policy Design 179
6.4 Methodology: the Regression Discontinuity Design 181
6.4.1 Estimation 182
6.4.2 Inclusion of covariates 185
6.5 Results 186
6.6 Robustness Checks 192
6.7 Conclusion 195

Summary of main findings 205

Bibliography 211

List of Tables

Table 1.1: Per-learner spending in public ordinary schools by province for selected years between 1998 and 2012 20
Table 1.2: Actual provincial allocation per learner against national targets, 2012/13 (Rand) 21
Table 1.3: Median departmental and total school NPNC spending per student (ZAR) 22

Table 2.1: Predictive performance across model parameters 51
Table 2.2: Summary of the relative contributions (%) of controls for boosted regression tree models of numeracy and literacy test scores 52
Table 2.3: Strongest two-way interactions for Grade 4 numeracy and literacy BRT models 58
Table 2.4: Most influential predictors across BRT models of numeracy score using sub-samples of the NSES (2008) data 62
Table 2.5: Most influential variables in BRT models of numeracy across the NSES Grade 5 (2009) and SACMEQ Grade 6 (2007) datasets 64
Table 2.6: Model performance of competing approaches using training and test data splits 65
Table 2.7: Variable importance across boosting and random forest models of numeracy and literacy 67

Table 3.1: Aggregate decomposition results 90
Table 3.2: Detailed aggregate decomposition results 93
Table 3.3: Sensitivity checks based on propensity score selection rules, alternative model specifications and matching procedures 95
Table 3.4: Detailed decomposition results for different model specifications 97

Table 4.1: National benchmarks for selected school and classroom resources 117
Table 4.2: Estimated sample and population treatment effects of attending an EAT school 130

Table 5.1: Cross-sectional regressions 153
Table 5.2: Correlated random effects models 154
Table 5.3: Correlated random effects models across sub-samples 156
Table 5.4: Correlated random effects models across different school sub-systems 159
Table 5.5: Returns to other teacher and classroom characteristics 163
Table 5.6: Correlated random error model results using the ST sample 165
Table 5.7: Student fixed effects estimation results 167

Table 6.1: Comparison of compulsory and non-compulsory tutorial attendance groups 188
Table 6.2: Regression results for tutorial attendance and performance 191
Table 6.3: Non-parametric results 192
Table 6.4: Sensitivity checks 195

Table of Figures

Figure 1.1: A basic systems model of school effectiveness 16
Figure 1.2: A simple context-led model for conceptualizing quality of education 28

Figure 2.1: Kernel densities of grade 4 numeracy scores in 2008, by former school department 49
Figure 2.2: Example tree structures and partial plots from a BRT model 50
Figure 2.3: Partial dependence plots for the nine most influential variables in the model for grade 4 numeracy 54
Figure 2.4: Partial dependence plots for the nine most influential variables in the model for grade 4 literacy 55
Figure 2.5: Joint partial dependence plots illustrating two-way interactions from Grade 4 numeracy model 60
Figure 2.6: RE-EM regression tree for Grade 4 numeracy 68

Figure 3.1: Reading score distributions, by school type 88

Figure 4.1: Reading test score distribution by school test language 122
Figure 4.2: Propensity score distribution across school test language 124
Figure 4.3: Difference in standardised means of student and home background covariates, pre- and post-reweighting 126
Figure 4.4: Difference in standardised means of school, classroom and teacher covariates, pre- and post-reweighting 127
Figure 4.5: Difference in standardised means of school, classroom and teacher covariates not included in coarsened exact matching 129
Figure 4.6: School treatment effect (SATO) as a function of balance in school level covariates 132

Figure 5.1: Student performance by school SES quintile 157
Figure 5.2: Teacher performance by school SES quintile 157
Figure 5.3: Returns to teacher knowledge by performance quintile and school wealth group 160

Chapter 1

Introduction

1.1 Background

More than two decades into democracy, South Africa remains a society divided. Despite its final dismantlement in 1994, the enduring remnants of apartheid are inescapably evident within the education system, where fault lines that are drawn by race, socio-economic class and geographical location continue to contribute to inequities in school quality and consequently, educational performance and attainment. Under the apartheid regime, the government allowed for separate and racially defined education departments,[1] each providing quite divergent types and qualities of education. Besides tangible deficits in resources,[2] schooling under the Bantu education system[3] also sought to indoctrinate conformity, rote learning and authoritarian management styles.

Despite concerted efforts to equalize expenditures per learner within the public education sector since 1994, the highly divided and unequal schooling system that was inherited from the apartheid regime has meant that many of the former black African schools that were entirely dysfunctional under apartheid remain dysfunctional today (Spaull, 2013). This is evidenced by high rates of dropout and grade repetition, underperformance and gross levels of teacher absenteeism amongst the poorer parts of the South African schooling system (Taylor, Muller & Vinjevold, 2003).

It is now commonly accepted that the average performance of South African students – both internationally and regionally low - masks a bimodal distribution of results; approximately 25% of students, most of whom come from affluent home backgrounds, attend high quality schools, whilst the remaining 75% of (predominantly poor and black African) students are found to attend low-quality and highly dysfunctional schools. This two-tier schooling system further translates itself into the labour market, where the latter group of students has little, if any, chance of furthering their studies past secondary school. And so it is that the low earnings potential linked to an inferior quality

[1] The institution of a racially sub-divided education system saw the creation of separate administrative departments for white schools (House of Assemblies (HOA)), Coloured schools (House of Representatives (HOR)), Indian schools (House of Delegates (HOD)), black African schools (Department of Education and Training(DET)) and each of the nine homelands.

[2] In 1986, students in white schools were subsidized R2 365 per capita; this is compared to R572 within the former Department of Education and Training (DET) schools. In 1992, this difference was still fourfold (Chisholm, Motala & Vally, 2003)

[3] The official system of education for black African South Africans.

of education that is itself linked to poor socio-economic status driven by poor labour market prospects perpetuates itself. It is therefore imperative that the quality of education, particularly that which is provided to the poorest of society, be improved if these cycles of entrenched poverty are to be broken.

No commonly accepted definition of quality exists, and defining quality in relation to education is especially difficult. Much of what has been understood by "quality" in education has sprung from Western episteme, with discourse largely dominated by human capital and human rights approaches (Tikly, 2011). In its 2005 Global Monitoring Report, UNESCO identified three education traditions associated with notions of education quality. These were termed behaviourist, humanist and critical, each with their own epistemological foundations that correspond to alternative education and development discourses (Yates, 2007). For example, human capital theory can be viewed as having an affinity to behaviourism where quality is evaluated through input-output models (learning as consequences), while a human rights approach can be linked to humanism where quality is evaluated based on process (learning as constructions).

In conceptualizing and understanding the role that quality education (or lack thereof) plays in South Africa, this thesis adopts a recently developed social justice framework (Tikly & Barrett, 2011). This new theoretical approach questions the assumptions and values inherent to the dominant approaches, as well as posits new understandings through drawing insights from social justice theory and the work of Amartya Sen and Martha Nussbaum in the area of human capabilities.[4] The social justice approach offers a synthesis of the human capital, human rights and critical approaches, allowing researchers and policymakers to consider and work on policy challenges within educational quality by drawing on the best of what is known from all the relevant discourses. As Sen argued: "we must go beyond the notion… after acknowledging its relevance and reach. The broadening that is needed is additional and cumulative, rather than being an alternative…" (cited by Robeyns, 2006:75). This brings to the fore the need to seek out new methodologies that compliment this synthesis, as well as better reflect the realities of stakeholders in education based in developing countries. It has become clear through international and comparative education studies that individual students and groups of students experience quality of education in different ways. Furthermore, many different barriers (such as gender, home language, socio-economic standing and rurality) work to prevent disadvantaged students from benefiting, or at a minimum accessing, good quality education.

[4] The capability-approach states that people should be afforded the freedom to achieve what Sen (1997) refers to as "functionings" (e.g. self-respect) that can be defined as "their real opportunities to do and be what they have reason to value" (Robeyns, 2009). Important contributions to the area of human capabilities specifically within the field of education have been made by Robeyns (2006), Walker (2006) and Unterhalter (2007).

This introductory chapter continues by first outlining the human capital and human rights frameworks that have dominated the educational quality literature. This will be followed by a discussion of the policy approaches that have been adopted within the South African schooling system since 1994 with emphasis placed on how these policies have been informed by the aforementioned human capital and human rights approaches. Following this, the social justice framework and its applicability to the South African context will be discussed. This chapter concludes with the research response through an outline of this thesis.

1.2 Dominant Approaches to Understanding Education Quality

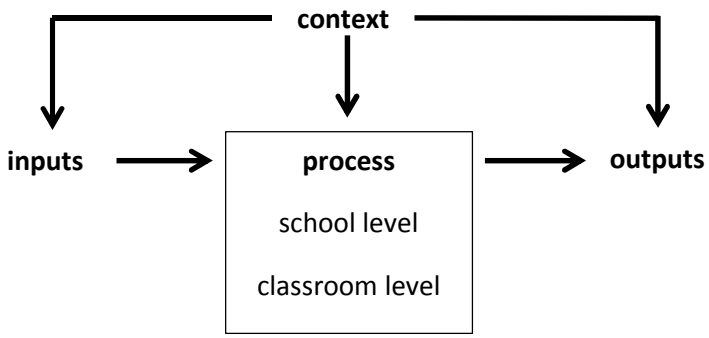
1.2.1 Human capital approach

This has been the dominant discourse in terms of the quality debate, as well as been influential in policy formulation. The human capital approach motivates investing in education given the positive contribution that it makes to development. Theodore Schultz and Gary Becker pioneered this conceptualization in their seminal work of the 1960s (Becker, 1962; Schultz, 1961) and has become well-established in economic theory. On a greater economy-wide level, greater human capital (skills and knowledge) should improve labour productivity and innovation as well as facilitate the transmission of new knowledge and technology. At the level of the individual, education improves individual productivity, ultimately leading to greater labour market employability and earnings potential (Mincer, 1974). These roles of education are what Robeyn (2006) refers to as *instrumental collective* and *instrumental personal* economic roles. The fact that human capital theory places people, as opposed to, for example, technical progress, at the centre of economic development is particularly important in contexts of high poverty and high inequality where even basic levels of literacy and/or numeracy can have significant effects for achieving minimum standards of living.

As the human capital approach does not in itself provide a framework for understanding educational quality (Tikly, 2011), school effectiveness frameworks have often been applied. The basic design of school effectiveness research typically adopts a linear input-output production function (systems) model in which school quality exists as the relationship between the teacher, classroom and school organizational environment and the student (Fuller, 1986). The school and classroom processes are generally viewed as something of a “black box”. The main task of school effectiveness research is to reveal the impact of relevant inputs - in the form of financial and material resources, teachers and pupil characteristics - on output; that is, break open the black box in order to show which educational processes or factors work. The research focus in school effectiveness studies can vary according to which factors or processes are believed to have originated the educational output; for example, the role of the school in creating equality of

opportunities in education and studies of education production functions are common research themes.

Figure 1.1: A basic systems model of school effectiveness



Source: Scheerens (1999)

In terms of policy intervention to raise the quality of education, human capital theory primarily advocates market-led solutions that are often grounded in rational choice theory. For example, Hanushek and Woessmann (2008) highlight three key areas that reform aimed at addressing educational quality should address: (i) creating greater choice and competition between schools; (ii) increase school autonomy including fiscal decentralization and local decision making; and (iii) greater accountability through the use of external examinations and benchmarking.

1.2.2 Human rights approach

While economists tend to think about education primarily in human capital terms and emphasize economic growth as the object of development, the human rights approach emphasizes the realization of fundamental human rights; it is interested in rights to education, rights in education and rights through education (Subrahmanian, 2002). These include the enactment of both negative rights (the right not to be abused) and positive rights (the right to use one's mother-tongue language), although in practice the former tends to be emphasised. At the level of policy, the right to education is probably most directly related to the "Education for All" movement, and, because education is seen as the right of every child, it is the duty of government to organize public resources so as to offer a quality education. Obligations derived from the right to education can be categorized as to make education *available* (ensure compulsory and free education), *accessible* (eliminate exclusion from education based on for example race, gender and language), *acceptable* (set minimum standards for education, implement a non-discriminatory curriculum and ensure that the entire education system conforms to human rights) and *adaptable* (Sandkull, 2005).

In opposition to the "black-box" treatment of classroom processes within the human capital approach, the human rights approach promotes learner-centred teaching and democratic school structures (Tikly, 2011), as well as the enhancement of social cohesion. One example of a framework that assumes a learner-centred view of education quality is adopted by the Global Campaign for Education (GCE) (2002) and is organized around five dimensions: what students bring to learning; environments; content; processes; and outcomes. Within a rights-based approach the non-economic instrumental roles of education can be realized. At an individual level, rights through education might be achieved as a result of the ability to speak with strangers in their languages, or the ability to work with technology enabling communication with people across the world. At a collective level, rights in education can, for example, allow children to learn to live in a more tolerant society.

1.2.3 Criticisms of the human capital and human rights approaches

Although the human capital and human rights approaches have provided the foundation for many of the policy initiatives in education, they are far from comprehensive and not without their limitations. The first issue with human capital theory is that it tends to ignore, or least downplay, the cultural, social and non-material dimensions of life (Robeyns, 2006: 72). Although the simple school effectiveness model depicted in figure 1.1 indicates contextual factors at all stages of the educational production process, the primary interest of the model is to explain the effectiveness or efficiency of the system, thereby reducing the role of contextual conditions to one that is secondary. Internal and external restrictions on learning are therefore not fully accounted for and their implications cannot be problematized and modelled. This links to a second issue with the human capital approach: it is problematic to assume that a linear relationship between student background characteristics, enabling educational inputs, processes and outputs exists. The inter-relationships are complex, multi-dimensional and vary with context.

A potentially hazardous result of using input-output model is that it can prescribe a one-size fits all approach to quality, thereby prescribing that the provision of a particular enabling input or the use of particular classroom process to emphasized which might only work for some students in certain schools. Limited resources at a government's disposal will then be directed to those factors or processes that yield the greatest return, as identified by the input-output model. Finally, education policies that are based on market-led solutions, as advocated by human capital theorists, can often exacerbate rather than reduce inequality in educational outcomes, which is in direct conflict with the development goals of the human capital approach which proposes reducing inequality through educational investment (Tikly, 2011).

With regards to the human rights discourse, one immediate issue with the approach is that its prescriptions feel rhetorical to the point of being blatantly obvious. Yet, despite the fact that most countries have extended the legal right to education to all children, this does not correspond to all children being present in schools; in many cases, children are attending schools where no teaching is taking place. As Subrahmanian argues: *"the haste to achieve 'education for all' has been interpreted in policy terms as [a] race of numbers, rather than a shift towards the creation of the kind of education system that can embrace diverse groups, and acknowledge and address economic constraints that limit participation in education*" (Subrahmanian, 2002: 2). When rights are pitched at the level of policy and legislation, this is precisely where it might end. There is the risk that once governments enact rights-based educational policy no further responsibilities or claims beyond fulfilling this obligation can be placed on them. Additionally, little notice might be paid to grass-roots level campaigns for quality education, the channels through which rights are to be effectively and precisely executed ignored (Robeyns, 2006). A final issue with the human rights approach is that it tends to be government-focused such that the state and not families and communities are held accountable for failing to provide children with access to good education.

1.3 Education Policy in South Africa Post-Apartheid

The twenty-years since 1994 have introduced a radically new historical era for education in South Africa. Anything that had been systematically linked to apartheid was abolished and replaced with new policies aimed at upturning prevalent inequalities, with the provision of universal, quality education a top priority. In all policy documents that have been produced in South Africa since 1994, for example, the White Paper of Education and Training (1995), the National Education Policy Act (1996) and the Culture of Learning, Teaching and Service (COLTS) campaign, quality and equality are emphasized. It is laudable that much of the progressivism has been concerned with achieving equitable education, particularly for those students who have been disadvantaged by public schools (Mouton, Louw & Strydom, 2012). However, the reality in many South African schools today reflects an alarming absence of both quality and equality. Before returning to this point, I will first highlight some of the primary policy strategies and reforms since 1994, as well as discuss a few of the transformation successes.

The policy stance that has been adopted since democratization is most neatly summarized in the preamble to the South African Schools Act (SASA) of 1996:

" [T]his country requires a new national system for schools which will redress past injustices in educational provision, provide an education of progressively high quality for all learners and in so doing lay a strong foundation for the development of all our people's talents and capabilities,

advance the democratic transformation of society, combat … all … forms of unfair discrimination and intolerance, contribute to the eradication of poverty and the economic well-being of society, protect and advance our diverse cultures and languages, uphold the rights of all learners, parents and educators, and promote their acceptance of responsibility for the organisation, governance and funding of schools in partnership with the State."

Educational transformation in South Africa has been premised on the achievement of the goals of access, redress, equity and quality, with schools expected to promote democracy as well as other freedoms (for example, the protection of culture and language). It is evident from the SASA (1996) that the policy approach to education has borrowed from the human capital and human rights discourses as well as the notion of human capabilities.

Some of the focal aspects of educational reform that address the abovementioned goals include: (i) equalising of public expenditure on education; (ii) the provision of free and compulsory education for 10 years; (iii) restructuring of school ownership, governance and finance; (iv) the introduction of new curricula; and (v) the establishment of new education management structures. Regarding point (v), the 19 racially defined departments under the apartheid regime were agglomerated into one national school system with the additional creation of nine provincial departments. Although the national department of education shares a concomitant role with the provincial departments for the provision of basic education,[5] the latter are responsible for the financing and management of schools within their respective province whilst the former is tasked with providing coherence of policy and philosophy (Chisholm, 2004).

1.3.1 Focus on Equity, Redress and Access

It can be argued that one of the most salient features of the South African schooling system is its entrenched structural inequality. However, it is clear that the immediate response by the post-apartheid government to equalize public expenditures across schools has resulted in a spending climate that has become equitable and even progressive. From columns 1 and 2 of table 1.1 there has been a notable improvement in the distribution of educational spending across provinces between 1998 and 2012. Spending per learner across provinces had reached almost equal distribution in 2012, with the highest public expenditure per learner approximately 18% higher than the lowest; this is compared to 75% in 1998/99.

[5] Basic education in South Africa covers early childhood development (ECD), primary schooling and secondary schooling.

However, increases in spending have not necessarily translated into real resources shift. Increases in spending have largely come about through rising teacher salaries, most recently occurring through the Occupation Specific Dispensation introduced in 2007, and have occurred particularly within the former disadvantaged school system (Van der Berg, 2007). From columns 3 and 4 of table 1.1 we can see that expenditures on personnel account for more than 90 percent of total education expenditures in most provinces. As a result, an average 8% of provincial education department's budgets are distributed for non-personnel expenditures. In the North West Province where 98% of education expenditure is allocated to personnel spending, the estimated per learner allocation for non-personnel non-capital (NPNC) inputs (such as learning and teaching support materials and school maintenance) is R175 as opposed to the average target of R814 (Financial and Fiscal Commission, 2014: 113). Contrastingly, the Kwa-Zulu Natal, Gauteng and Western Cape provinces have a per-learner NPNC allocation that is 65-100% larger than prescribed. This can in part be explained by the ability of (mainly wealthy) schools to raise additional private funds through school fees; I will return to this point later on.

Table 1.1: Per-learner spending in public ordinary schools by province for selected years between 1998 and 2012

Province	ZAR/learner in public schools 1998/99	ZAR/learner in public schools 2001/02	ZAR/learner in public schools 2012	Proportion personnel expenditure (estimate) 2002/03	Proportion personnel expenditure 2012
Eastern Cape	2 884	3 878	10 639	94	90
Free State	3 291	4 509	11 751	89	92
Gauteng	4 206	5 031	10 469	86	86
Kwazulu-Natal	2 575	3 481	10 349	92	87
Limpopo	3 165	3 720	10 495	92	93
Mpumalanga	2 851	3 725	10 708	93	93
Northern Cape	4 526	5 256	10 697	84	94
North West	3 374	4 496	9 886	92	98
Western Cape	4 171	4 875	10 506	88	90
National average	3 449	4 330	10 533	90.8	90.2

Source: National Treasury (2003) Intergovernmental Fiscal Review; Financial and Fiscal Commission (2014), Submission for the Division of Revenue

According to the Norms and Standards for School Funding (NNSSF) (2006), the distribution of non-personnel funding within provinces is meant to be pro-poor. Schools are ranked and placed into poverty quintiles based on (i) the poverty of the school community and (ii) school conditions, with the result that resources be allocated based on this school poverty index. The poorest schools (quintiles 1, 2 and 3) are classified as fee-free schools and are meant to receive 80% of the available

NPNC funding.[6] The minimum no-fee threshold spending per learner is R926. From table 1.2 we can see that whilst most provinces meet prescribed spending levels (or at least the minimum threshold); some provinces are underfunding the poorest schools whilst others are overfunding the wealthiest schools. This not only suggests an inequitable distribution of resources among provinces, but also poor fiscal management by provinces. Insufficient capacity within provincial and district level management to process schools' requests for goods and services have led to late delivery as well as late financial transfers (Taylor, 2010: 22).

Table 1.2: Actual provincial allocation per learner against national targets, 2012/13 (Rand)

Quintile	National target	EC	FS	GT	KZN	LP	MP	NC	NW	WC
1	1010	926	1010	1010	932	808	1010	1010	1010	1012
2	1010	926	1010	1010	932	740	1010	926	1010	1011
3	1010	926	1010	1010	932	740	1010	926	1010	1011
4	505	505	505	505	505	505	505	505	606	548
5	174	174	174	240	505	174	138	174	174	250

Source: Financial and Fiscal Commission, Submission for the Division of Revenue (2014)

Despite these indications of lags in creating fiscal equity, there have been impressive improvements in school infrastructure over the last two decades. The numbers of schools with access to electricity, water and sanitation have nearly doubled (OECD, 2008), although infrastructural backlogs still persist with regards to access to libraries, computers and science laboratories. According to the Department of Basic Education's National Education Infrastructure Management System (NEIMS) report of 2011, approximately three-quarters of schools did not have libraries or computer laboratories, and amongst those who did have these facilities only 7% were fully stocked. Policy makes no prescriptions with regards to how much provincial education departments must budget for school infrastructure, or to which schools it should be allocated; rather, it simply states that funding should favour "redress and equity".

Access to basic education in South Africa has reached almost universal levels. At least 99% of children enter formal education with dropouts being very low until Grade 7 (end of primary school education). With the roll-out of Grade R within the public education system, the numbers of Grade 1 learners who attended pre-primary increased from 242 000 to 768 000 between 2001 and 2012 (ReSEP report prepared for DBE). This corresponds to approximately 75% of all Grade 1 learners. Over the same period, the proportions of children attaining at least a Grade 9 have risen from 76% to 86% (Spaull, 2012). However, greater access to schooling has not translated into qualitative

[6] This used to be the poorest 40% of schools who were allocated 60% of the available NPNC funding.

improvements in schooling outcomes as issues of redress and equity in school play themselves out through school choice and admission policies (Spreen & Vally, 2006). Section 247 of the interim Constitution and Section 21 of the SASA (1996) afforded considerable powers to school governing bodies (SGBs) whereby local communities were made progressively more responsible for raising and spending privately acquired funds, typically through user-fees.[7] The rationale behind this was that user fees would supplement public spending in communities that could afford it whilst simultaneously allowing government to redistribute funds to the poorest schools.

It is now argued that quality has been reduced to what can be raised through school fees, with good quality education in South Africa linked to the likelihood of residents in the local community being able to afford investments in schooling (Yamauchi, 2011). User fees have allowed for the maintenance of higher quality facilities in Section 21 schools with the subsequent movement of children whose parents are able to pay high user fees into better resourced schools. The result: a yawning gap in resources between rich and poor schools on the one hand and a yawning gap in performance between rich and poor students on the other. Private spending in the form of school-fees (and to a lesser degree fund-raising) changes the picture of equalization to one of substantial divergence within the public sector. Table 1.3 shows median school NPNC spending per student (made up of school fees and government transfers) plus departmental spending by school poverty quintile in 2009. Non-personnel departmental spending norms aimed for approximately 6 times the expenditure in quintile 1 than in quintile 5 schools; from the first row of table 1.3 we can see that the reality was closer to 3 times. Once private spending from school fees is included (second row of table 1.3) we can see that spending per student in quintile 5 schools is roughly 3 times that in quintile 1. This is in exact reverse to what the policy intention of creating equity in expenditures hopes to achieve for promoting redress amongst the disadvantaged student population.

Table 1.3: Median departmental and total school NPNC spending per student (ZAR)

	Departmental spending	Total school spending
Quintile 1	711	981
Quintile 2	711	944
Quintile 3	481	1062
Quintile 4	474	1105
Quintile 5	228	2829
Total	591	1673

Source: DBE (2009: 47)

[7] In 1990 most white public schools were granted the right to appoint teachers, decide on school fees and impose admission policies. These schools are referred to as Model-C schools. This enabled the preservation of a privileged white public school system in the wake of the collapse of apartheid.

The notion of a "bimodal distribution" within the South African education system has become commonplace within educational research, revealing itself to be impervious to the grade or subject being analysed. Whether the sample is split by school wealth, school language of learning and teaching or former education department, the performance of students attending wealthy/English-Afrikaans/former white schools can be as much as 2 standard deviations higher than students attending poor/African language/former DET/Homeland schools (see for example Spaull, 2013; Taylor, 2011; Shepherd, 2011).

1.3.2 Teacher Interventions and Curriculum Reform

As with expenditure, the uneven and racially hierarchical provision of educators that had been created under apartheid required urgent attention from 1994. Teacher employment was brought under a single Act of Parliament and a new teaching post-distribution (provisioning) system negotiated that was based on teacher-student ratios, subject fields and language of instruction. This implied that schools catering to a large number of students and/or offering more diverse curricula were allocated more posts. In 2002, this model was revised to take into account school poverty quintile such that provinces are permitted to retain a maximum of 5% of available posts to be allocated as "redress" posts, with 80% allocated to quintile 1, 2 and 3 schools (Financial and Fiscal Commission, 2014: 118). This post-provisioning process has unintentionally favoured more "affluent", mainly former white, schools where subject choices are more varied. This fact combined with the private funds generated through user-fees has meant the maintenance of staff numbers and small class sizes within these schools. Between the years 1996 and 2000, teachers paid from state coffers decreased by close to 24 000 while SBG-paid teachers increased by 19 000 (Spreen & Vally, 2006). As the 1998 Norms and Standards for School Funding mentions: "Ironically, given the emphasis on redress and equity, the funding provisions of the Act appear to have worked thus far to the advantage of public schools patronized by middle-class and wealthy parents … since 1996, when such schools have been required to down-size their staff establishments, many have been able to recruit additional staff on governing body contracts, paid for by the school fund" (amended National Norms and Standards for School Funding, 2006: 10).

In terms of the training and education of teachers, the government has successfully managed to significantly reduce the numbers of unqualified and under-qualified teachers in the system, although this has mainly occurred through in-service upgrading programmes. 36% of educators were considered un- or under-qualified in 1994; this proportion declined to 8.3% in 2004. Despite this, the majority of teachers continue to be unequipped in terms of subject knowledge and pedagogical skill. This is most likely due to the fact that most teachers currently serving as educators

in the public school system were trained before 1994 when teacher demand requirements of the whole country were largely disregarded (OECD, 2008: 83). Teacher recruitment, training, deployment and motivation are particularly challenging issues when education systems expand rapidly (Tikly & Barrett, 2011: 9).

A further major component of education policy post-1994 has been curriculum reform as a driver for quality. *Curriculum 2005* was launched with the purpose of nation building and fostering inclusive education (Taylor, 2010: 24). The philosophy of Outcomes Based Education (OBE) was believed to support this notion of a rights-based national curriculum: "… our education system and its curriculum express our idea of ourselves as a society and our vision as to how we see the new form of society being realized through our children and learners" (Revised National Curriculum Statement (RNCS), 2002: 1). Notwithstanding its broad-based support, fundamental problems with *Curriculum 2005* soon began to reveal themselves as OBE became embroiled with the everyday realities of South African classrooms. Despite OBE being aimed at empowering teachers it emerged as too complex and deficient in directive. Lack of clarity of design, language and terminology ("the curriculum is and will be differently interpreted and enacted in diverse contexts" (Department of Basic Education, 2002)) combined with a lack of teacher training and support further limited its successful implementation. Qualitative research that has been sensitive to the viewpoints and lived realities of teachers' practices have suggested that some teachers opt to facilitate learner participation in ways that address the broader socio-economic contexts of their classrooms (Barrett, 2007; Mtika & Gates, 2010). *Curriculum 2005* was simplified in the RNCS with more prominence given to basic skills, content knowledge and teacher support. From 2012 the curriculum has been combined into a single document, the National Curriculum Statement (NCS), for Grades R to 12. Building on the previous curricula, the NCS aims to provide a clearer specification of what is to be taught and learnt.

1.3.3 Policy Lessons

With the establishment of new management structures, it was believed that the national policy vision for school practice would somehow trickle down the provincial and district layers. The achievement of educational quality through legislation and policy anticipated a fairly smooth process of increasing the system's capacity and a redistribution of human, physical and material resources. Yet, in spite of a nationally agreed framework, every stage of policy implementation has presented with greater or lesser degrees of conflict. There appears to be a great disconnection between the policy norms and standards that are set at the national level and how these are understood and enacted at the provincial level. Furthermore, fiscal and capacity constraints at provincial level have

meant that provinces are struggling to keep within budget whilst simultaneously meeting the urgency of delivering visible reform.

A clear example of how policy reform aimed at creating equity within the public school system has potentially reinforced inequality in educational opportunities and outcomes is the semi-privatization of public schools through the extension of financing and governance provisions to SGBs (Lemon, 1999). Allowing all schools to raise funds is perhaps the most direct means of addressing the budgetary limitations of government as well as limiting the flight of white children out of the public school system (Selod & Zenou, 2003). However, this reform has ignored (or denied) the existence of a spatially determined distribution of income and population groups that preserve inter-racial and -socio-economic diversity in access to good education as the best schools continue to be located within selected areas. Financial constraints pose a real threat to poor children in accessing a good quality education (Dieltiens & Meny-Gibert, 2012). Furthermore, despite the implementation of the 1996 SASA, the private schooling sector has burgeoned not only as a result of higher demand amongst middle-class (mainly white) students but also amongst disadvantaged communities where low-fee private education is becoming increasingly available and a financially viable alternative to public schooling.

The policy approach since 1994 has illustrated that although transformation is necessary, it is not sufficient to ensure *real* educational transformation. One of the key difficulties faced by policy makers is the need to shift from a positivist view to a more systemic way of understanding schools and the process of change. It could be argued that the post-apartheid government went for second order change; that is, fundamentally changing the way in which schools are structured and roles are defined, without also developing the capacity of the education system to make and implement good policy. In addition, whilst transformation has emphasized the use of legislation and regulatory frameworks to put systems in place, pedagogy and the actual process of teaching and learning has been until recently largely ignored. Successful second order change within education entails: (1) a fundamental change in ideas about and actions towards student achievement; (2) instructional enhancement that is attentive to pedagogy; and (3) collaborative support that instils a culture of widespread partnership (Baker, 1998). The redistribution of resources is insufficient in itself; it should lead to a redistribution of the conditions of learning such that equity in learning achievement is possible (Crouch, 1996). Heneveld (1994) suggests that the processes within schools and classrooms that contribute significantly to school effectiveness are to a large degree independent of policy.

Elmore (1996) points out that the 'core' activities of educational practice are very hard to change, especially through policy action. These activities can be defined as: how teachers

understand the nature of knowledge; how teachers understand the students' role in learning; how ideas about knowledge and learning are put into practice in teaching and classwork; and the structural arrangements that support teaching and learning (for example, physical layout of classrooms and processes for assessing student learning) (Christie, 2008: 151). The (relatively speaking) easy structural changes that can be made, for example, in school governance and financing can have significant symbolic value, but do not any actual change to teaching and learning. Christie (2008: 152) argues that the same can potentially be said of elaborate reporting and accountability procedures which give the appearance of tackling quality issues, but do not bring about any purposeful change in the conditions of schools and classrooms.

School effectiveness and school improvement research in South Africa (and elsewhere) has shown that the answer, in very broad terms, to the question "what will make a difference to the learning outcomes of different students at school?" are what students bring with them to school in terms of their home backgrounds, which schools they attend, how well their schools function, how effective their teachers are and what happens inside the classroom (Christie, 2008: 164). What we require is a better understanding of the school (including student, teacher and classroom) factors that together, not in isolation, form the social setting that conditions how teaching and learning takes place. Understanding the interaction and linkages between poverty indicators, level of schools resources and school outcomes can provide a more holistic understanding of the barriers facing different groups in accessing a good quality education. This requires questioning the assumptions implicit to our current understanding of quality as well as the use of new and innovative methodologies that can reflect, as far as possible, the realities of South African classrooms and learners.

1.4 Social Justice Approach to Education Quality

Fraser (2009) highlights three dimensions of social justice (redistribution, recognition and participation) that are each related to institutional and structural barriers (economic, cultural and political) that impinge on the realization of human capabilities. These dimensions can be identified as three inter-related principles that provide a benchmark against which an education system could be assessed: (1) education should be inclusive; (2) education must be relevant; and (3) education should be democratic (Tikly, 2011). Social justice is generally understood as being primarily concerned with redistributive justice. In the context of education this implies access to quality education and the potential outcomes that arise from this. The focus of this thesis is primarily the dimensions of redistributive justice and inclusion within the South African schooling system, although I will briefly discuss the other two dimensions.

Justice through recognition implies the identification and acknowledgement of the claims of historically marginalized groups and requires equal respect regardless of race, gender, religion etc. be extended to all participants. This is achieved through the Constitution and the Bill of Rights. Participatory justice, whereby individuals and groups have rights to make claims for social justice and actively participate in decision-making, is a prerequisite for realizing the dimensions of recognition and redistribution (Tikly & Barrett, 2011). The opportunity to participate is seen as an essential indicator of how democratic a state is. The establishment of SGBs hoped to bring those closest to the schools into the decision-making process and through doing so deepen the educational experience. However, "it is not enough to be included in the decision-making process; one also needs to be able to influence the process and the decision" (Dieltiens, Chaka & Mbokazi, 2007: 13). Placing any kind of expectation on SGBs to transform schools should be measured against the ability of SGBs to deliberate issues in any kind of participatory or democratic way.

What is clear from the discussion thus far is that a narrow focus on simply the distribution of resources (expenditures) and a fixation with simple equality can obscure the real issues at stake in the pursuit of social justice (Pendlebury & Enslin, 2004: 1). A principal issue related to redistribution is the absence of a clearly formulated definition of quality, sometimes limiting its achievement as simply an increase in outcomes. This has reinforced the tendency to observe the educational process as a "black box" whereby teaching and learning processes are neglected. From a social justice perspective, inclusion is concerned with the access that different students have to a good quality education and the opportunities for achieving anticipated outcomes (Tikly & Barrett, 2011: 9).

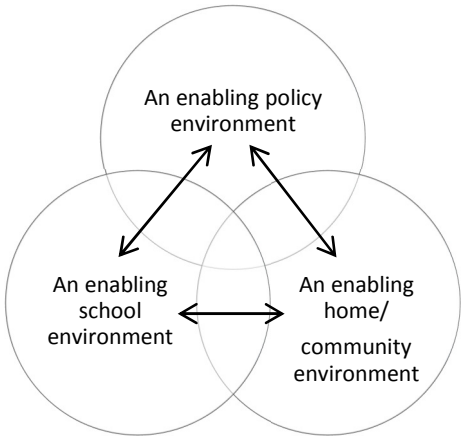
In order to better target resources and interventions in education, a refined understanding of the different kinds and levels of resources required by different groups of students is needed. School effectiveness studies consistently point towards the importance of textbooks and other learning materials for raising student outcomes, but more so than the simple provision of learning and teaching support materials is that they be dependent on and customized to the pedagogic practices, professional values and language proficiencies of teachers. Teacher quality and pedagogy have increasingly become central to the quality debate. Ensuring inclusion requires continuous monitoring of quality and the disaggregation of student outcomes to reveal who are disadvantaged as well as the barriers that operate to prevent students from accessing resources and converting them into capabilities. The recognition of a bimodal distribution of performance in South Africa has therefore been a positive step toward beginning to understand the nature of inequality in educational opportunities and outcomes.

An argument can therefore be made for making context implicit to a definition of quality education. This is reflected in an emerging framework founded on social justice principles by Tikly

(2011) that has been adopted by the EdQual[8] programme and expressly conceptualizes education quality in low-income countries. Specifically, a good quality education develops from the interaction between three overlapping environments: policy, the school and the home/community environments (see figure 1.2). Unlike the traditional input-output model, this framework highlights the importance of accompanying processes within each environment that result in the conversion of schooling inputs into outcomes. Furthermore, it does not limit the model to be linear, but rather identifies the achievement of schooling outcomes through a blend of inputs and processes and an interaction between environments. Creating a good quality education involves paying attention to the overlaps and ensuring that enabling inputs and processes work to close the gaps that exist between the environments (Tikly, 2011: 11).

The "implementation gap" between legislation and policy set at the national level and schools could be reduced by engaging with the experiences and perspectives of teachers and school principals, providing initial and continuing professional development and providing support to schools and teachers in implementing change. The "expectations gap" between educational outcomes and the expectations of parents and communities could be addressed through encouraging active participation in national debates and developing greater accountability within the system. Finally, the "learning gap" that exists between what takes place in schools and what is required of the home/community could be closed through working with parents to create a home environment that facilitates learning outside of school and providing school feeding programmes.

Figure 1.2: A simple context-led model for conceptualizing quality of education



Source: Tikly (2011: 11)

[8] Implementing Education Quality in Low-Income Countries

1.5 Thesis Outline

The recent availability of a number of rich nationally representative datasets has meant a resurgence of research into educational outcomes in South Africa. Internationally and regionally, South Africa has participated in three major cross-national comparisons of primary and secondary school student achievement, namely: the Southern and Eastern African Consortium for the Monitoring of Educational Quality (SACMEQ) surveys, Trends in Mathematics and Science Survey (TIMSS) and the Progress in Reading and Literacy Survey (PIRLS). At a national level, standardized testing programs have included the Systemic Evaluations of 2001 and 2007, the National School Effectiveness Study and most recently, the Annual National Assessments. All of these datasets have been analysed by academic researchers, policy- makers and educational NGO's yielding a considerable amount of insight into the performance of South African students, and the generative mechanisms behind that performance. Much of the existing findings speak to the types of enabling processes and inputs identified by the quality framework depicted in figure 1.2 (these studies and findings will be referred to throughout this thesis). The research conducted in this thesis aims to add to the current findings and literature through recognizing the complex and multi-dimensional nature of the issues relating to the quality of education in South Africa, in particular as it relates to and impacts on disadvantaged students. This implies going beyond the standard quantitative techniques (e.g. education production functions) to recognize the disproportionate impact of relevant variables on different groups of students.

The non-experimental nature of the collected data has meant that the majority of existing studies cannot infer causality and therefore only report partial correlations. Whilst descriptive assessments of the associations between schooling inputs and processes and student outcomes are valuable additions to the narrative of the South African schooling system, policy conclusions from causal evidence are sounder. Dealing with unobservable heterogeneity is fundamental to economic science. The availability of panel data is one way of coping with this issue, but in the absence of this type of data the researcher is forced to look for alternative methods. This thesis is therefore concerned with not only finding new and innovative ways to model and analyse the schooling process in South Africa, but also attempts to apply techniques that deal with the issues of non-random selection and unobservable heterogeneity so as to strengthen the case for making causal inference.

To address the current gaps in the research, I apply five distinct empirical methodologies: (i) boosted regression tree analysis to model grade 4 mathematics and reading performance within former DET and Homeland schools; (ii) parametric propensity score reweighting decomposition of

reading test scores across historically disadvantaged and historically advantaged schools; (iii) non-parametric overlap balance reweighting decomposition of reading test scores across historically disadvantaged and historically advantaged schools; (iv) within-student across-subject analysis of the impact of teacher knowledge on grade 6 performance; (v) regression discontinuity design analysis of the effect of a compulsory tutorial programme on first-year student performance. The remainder of this introductory chapter describes the five essays in Chapters 2 to 6. Chapter 7 provides a summary of the core findings, policy implications and guidance for future research.

Chapter 2: Tree of knowledge: A nonparametric approach to modelling primary school outcomes in South Africa

Chapter 2 of this thesis provides a relatively under-utilized methodology for modelling outcomes in the economic sciences, namely, regression tree analysis. Creemers and Kyriakides (2006) make a proposal for the use of dynamic models in educational effectiveness research (EER) that stems from three main criticisms of the existing school effectiveness research. First, the research evidence around certain classroom and teacher factors has been contradictory; for example, teacher subject knowledge rarely correlates strongly with student achievement. This may be related to the fact that the true relationship between teacher knowledge and student performance is curvilinear (Monk, 1994). Therefore, a dynamic model of EER should be based on the assumption that the relation of some effectiveness factors with achievement may be curvilinear. Second, EER models should take into account that effectiveness factors on the same or different levels (school, classroom, and home) can influence one another. Therefore, an approach to modelling schooling effectiveness should be able to reveal optimal combinations of factors that make teachers and schools effective. Finally, effectiveness factors should be considered as multidimensional constructs. Regression tree analysis allows us to address the first and second issue.

The chapter begins by making a general case for the use of flexible machine learning approaches for modelling education production as they allow for more complex response surfaces that are frequently observed in distributional data. Rather than relying on the traditional linear input-output model of education, the approach adopted here uses an algorithm to *learn* the relationship between test performance and its determinants, allowing for nonlinear relationships to be fitted between covariates and the dependent variable without having to specify any functional relationship/s. The analysis is restricted to a sample of former DET and homeland schools as primary interest is in understanding the mechanisms through which effective teaching and learning is created amongst the primarily disadvantaged subset of South African schools. The National School Effectiveness Survey (NSES) that identifies the former school department is employed. Findings

suggest that the maximum availability and use of time-on-task and opportunity to learn (coordination in curriculum and instruction) are salient contributors to learning outcomes. These classroom and teacher level factors combine with other factors at the same level (e.g. teacher experience and test scores) as well as home background factors of the students to produce augmented reading and mathematics performance.

Chapter 3: A question of efficiency: decomposing South African reading test scores using PIRLS 2006

This chapter aims to shed light on the source/s of discrepancy in performance between former black Africa /homeland schools and former 'advantaged'[9] schools, and whether the discrepancy comes as a result of differences in school quality[10] or access to a lower level of (quality) resources. The 2006 Progress in International Reading Literacy Study (PIRLS) that captures Grade 5 performance in reading is used to decompose the performance gap between those schools that tested in English or Afrikaans (as a proxy for the former advantaged school system) and those schools that tested in an African language (as a proxy for the former disadvantaged black African school system).

Botezat and Seiberlich (2013) employ a semiparametric approach for decomposing performance gaps in Eastern European countries. Their construction of a counterfactual mean using propensity score matching allows assessment of the extent to which differences in student and home background characteristics contribute to explaining the observable gaps in school performance (explained gap), with the remaining gap due to differences in schooling systems (unexplained gap). It is posited that constructing the unexplained gap in this way is more representative of the average treatment effect of attending a school within a particular school system. Unlike Botezat and Sieberlich, the analysis of this chapter adopts the reweighting approach of DiNardo (2002) and DiNardo, Fortin, and Lemieux (1996) to construct the counterfactual of interest. This approach allows the unexplained performance gap to be separated into two "treatments" of attending a different school type. The first of these is the effect of attending a school within a school system that offers higher returns to educational inputs, or to a school efficiency gap. The second component of the unexplained gap is due to differences in the distribution of school resources across the two school systems, or to a school resource gap. In this chapter I propose that these two components of the unexplained gap provide education policy with two different tools for assessing how the performance gap between two students attending schools within different school sub-systems might be closed.

[9] Here advantaged is meant to imply former white, coloured and Indian schools.
[10] School quality is defined as the extent to which a school and its constituent parts (teachers, management, culture and infrastructure) improve a student's learning.

The findings of this chapter indicate that home background factors play a significant role in explaining the test score gap, accounting for roughly 60% of the average test score gap between the two school systems. A further 14-35% (depending on whether or not school SES is included) of the average test score gap is accounted for by differences in observable school level inputs. The insignificant contribution of differences in teacher and classroom variables to the school resource gap provides evidence that the distribution of these factors is relatively balanced across former departments. This is, however, not to say that the quality of teaching and classroom processes are equal across sub-systems. Quality differentials as captured by the school efficiency gap are estimated to account for 16% of the average performance gap. Overall, the decomposition results estimated here predict that successfully addressing inequalities in the distribution of school resources (or processes) that augment performance whilst simultaneously addressing inequalities in school effectiveness or quality may as much as halve the average performance gap between the two former school departments.

Chapter 4: Balancing Act: A semi-parametric approach for determining the local treatment effect of school type

The analysis of this chapter follows on from that of chapter 3. The standard approach to assessing the effect of the type of school attended on student performance would be to imagine that the treatment assignment operates on students. Keele (2012) puts forward an argument that when interest lies in school effects, the hypothetical experiment would be one that focuses on assignment at the group (school) rather than the individual level. Taking treatment assignment to have occurred at the group level implies that covariate balance first be achieved at the school level before matching on student covariates.

The analysis of this chapter recognises that differences in the distribution of certain school resources may result post-treatment (such as factors related to good governance including greater teacher job satisfaction and lower teacher absenteeism) whilst others are likely to result pre-treatment (such as class size which might be policy mandated). Matching on all school resource variables, both post- and pre-treatment, would dramatically limit the comparative school samples. The analysis of this chapter proposes that in finding suitable comparator groups across the former disadvantaged and advantaged school systems, covariate balance be achieved on (i) pre-treatment school resources and (ii) students. As a result, the treatment effect will partly be a function of differences in the distribution of post-treatment school resources and partly a function of differences in school effectiveness across the two systems. Coarsened exact matching and overlap balancing weights are applied to the 2011 PIRLS dataset of Grade 4 reading scores. As with chapter

3, the treatment effect of attending a former advantaged school is estimated by comparing performance of student attending schools that tested in English or Afrikaans (as a proxy for the former advantaged school system) and those schools that tested in an African language (as a proxy for the former disadvantaged black African school system).

Achieving balance on student home background and pre-treatment school factors leads to an estimated treatment effect of attending an English/Afrikaans testing school that is equal to roughly 12-16 months of learning. The treatment effect is further shown to be a function of imbalances in school level factors. This speaks to the unequal distribution of school resources (such as teacher quality and teacher-student ratios) that is linked to the availability of private spending within Model C schools. Matching on school SES (as representative of the average wealth of the school) more than halves the size of the treatment effect. The methodological contribution of this chapter further indicates that, conditional on the ignorability assumption being satisfied, regression analysis serves as a viable alternative to matching and propensity reweighting estimators for estimating treatment effects.

Chapter 5: Learn to teach, teach to learn: A within-pupil across-subject approach to estimating the impact of teacher subject knowledge on South African grade 6 performance

This chapter investigates the role that teacher subject knowledge plays in determining student performance. One of the important challenges facing studies attempting to estimate the causal effect of teacher characteristics on student performance is the non-random sorting and selection of students and teachers into classrooms and schools. This issue may be addressed through the use of student and teacher fixed effects, although this requires the availability of longitudinal datasets that captures teacher subject knowledge. Given a lack of such data, this chapter makes use of a within-pupil between-subject methodology, namely a correlated random errors model, used by Metzler and Woessmann (2012) to estimate the effect of teacher subject content knowledge on grade 6 student test scores in South Africa. This methodology is an extension of the first differencing (fixed effects) technique proposed by Dee (2005, 2007) that has been applied quite extensively to eliminate bias from unobserved non-subject-specific student characteristics in order to identify the impact of various teacher and classroom factors. I further restrict the sample to students who are taught by the same teacher in the two subjects in order to correct for potential bias due to teacher unobservables.

Accounting for selection biases, teacher knowledge is estimated to have no significant effect on student outcomes. This is similar to the findings of Carnoy and Chisholm (2008) and Carnoy and Arends (2012) who find no significant effect of teacher content knowledge on student gains in

mathematics. However, this may mask differences in impact across student sub-groups. Separation of the sample by school wealth indicates that the impact of teacher knowledge is not homogenous across the South African education system. A significant positive non-linear effect of teacher subject knowledge is estimated for the wealthiest quintile of schools, whilst no significant effect of teacher knowledge is estimated for the poorest four school wealth quintiles. Teacher education is additionally estimated to have significant and large effects for student outcomes in wealthier schools, though this may be driven by a positive relationship to teacher unobservables. The same may be true of the large and highly significant effect size of young and inexperienced teachers in poor schools, which may signal an improvement in the training of those that have most recently entered the teaching profession.

A final finding of the analysis in this chapter suggests that teacher subject knowledge is positively related to teacher unobservable quality in the wealthiest 20% of schools; this is what we would expect. On the other hand, teacher subject knowledge appears to be negatively correlated to teacher (and school) unobservables in the poorest schools. This may be due to a lack of enabling factors contributing to effective teaching such as high quality training, pedagogical skill and opportunity to teach that are more present in wealthier schools. It may also suggest a correlation with factors that hinder the transmission of knowledge to students such as mismanagement, poor instructional leadership and poor teacher collaboration.

Chapter 6: Compulsory tutorial programmes and performance in undergraduate microeconomics: A regression discontinuity design (with Volker Schöer)

Although the research question and focus of chapter 6 does not appear to fit in directly with the remaining chapters, it poses a question that contributes to the overarching framework of this thesis; that is, can we identify potential interventions that can contribute to more effective learning, whether this is at the level of basic or higher education. Dropout rates amongst undergraduate students in South African universities are high, which comes with high financial and social costs. As with basic education, higher education departments are under constant strain to maintain quality whilst improving cost effectiveness of service provision. Minority student groups, which in South Africa are primarily disadvantaged students, may benefit disproportionately from a 'deeper' approach to learning (Entwistle, Thompson and Tait, 1992). There are therefore both practical and ethical reasons for the move towards adopting peer tutoring as part of the learning support structure in higher education.

The tutorial programme studied in this chapter was initiated by the Economics Department

of Stellenbosch University in 2009. Attendance of these tutorials was made mandatory for students that obtained below 50 percent in their early assessment test. Students who achieved at least 50 percent in the first test were still permitted to attend tutorials on a voluntary basis. The 2010 class cohort is used for analysis purposes given the stricter enforcement of the policy. The specific design of this policy presents an opportunity to directly assess the impact of tutorial attendance on academic performance through the use of regression discontinuity design.

The analysis of this chapter makes use of both parametric and non-parametric models for estimating the local treatment effect of attending the tutorials. Two-stage instrumental variable regression results indicate a significant 0.1 standard deviation increase in final exam performance for a 10 percent increase in tutorial attendance. Quantitatively similar impacts are found using local linear polynomial regression, although the results are sensitive to choice of bandwidth and specification of the control function. Robustness checks indicate that the results are fairly insensitive to the inclusion of the other covariates. However, the exclusion of the best performing compulsory students who were permitted to leave the programme decreases the treatment effect. This raises the concern that the result may be biased by unobservable factors such as motivation and effort that are not exogenous to the tutorial policy.

Chapter 2

Tree of knowledge: A nonparametric approach to modelling primary school outcomes in South Africa

This paper introduces a flexible non-parametric technique for modelling school effectiveness within the former disadvantaged school department in South Africa. Specifically, a boosted regression tree analysis is employed that allows for curvilinear associations between schooling factors and student outcomes, as well as interactions between schooling inputs, to be modelled. Results indicate that teacher inputs and classroom processes that allow for the availability and maximum use of time-on-task and opportunity to learn combine with home background characteristics to produce augmented test scores. The findings are robust to the use of sub-samples of the overall data and alternative datasets.

2.1 Introduction

Despite concerted efforts to equalise the distribution of school resources in the South African education system over the past two decades, a large portion of the system still fails to provide the quality of education needed to facilitate economic growth. International[11,] regional[12] and national[13] comparisons of South African student performance on standardised tests with both developed and much poorer countries continually highlight the generally weak performance of the South African basic education system.[14] Research indicates that the problem lies in the dismal performance of the historically disadvantaged, chiefly black, schools (Van der Berg, Wood & Le Roux, 2002: 305), with recent studies further indicating significant effects of attending a former advantaged, predominantly white school (Coetzee, 2014; Shepherd, 2013).[15]

[11] Trends in Mathematics and Science Survey (TIMSS) testing of Grade 8 students in 2003 and 2011, as well as the Progress in Reading and Literacy Survey (PIRLS) testing of Grade 4 and 5 students in 2006 and 2011.
[12] South African Consortium of Educational Quality (SACMEQ) testing of Grade 6 student in reading and numeracy in 2000 and 2007.
[13] Systemic Evaluations of Grade 3 students in 2001 and 2007, National School Effectiveness Study (NSES) testing a panel of students in Grades 3-5 from 2007-2009, and most recently, the Annual National Assessments (ANA) testing Grades 1-6 and Grade 9 in 2011 and 2012.
[14] Basic education refers to primary and secondary schooling running from Grade R through Grade 12.
[15] As referred it in the introductory chapter, The department for white schools was the House of Assemblies (HOA), for coloured schools it was the House of Representatives (HOR), Indian schools were administered by the House of

The substantial heterogeneity in the quality of schools available to students has emphasised the role that school choice plays in South Africa (Yamauchi, 2011). The geographic and socio-economic constraints faced by a significant number of predominantly black African households imply that many children have no other option but to attend a historically black (DET) or homeland school. Given that the vast majority (in excess of 80 percent) of South African schools fall within this group, I would argue that an enhanced understanding of the factors or processes that positively affect schooling performance within this schooling sub-system is required.

Attempts to understand the generative mechanisms behind student and school performance in South Africa have commonly adopted education production function type analysis. Quantities of measured schooling inputs are mapped, usually linearly, to a relevant measure of schooling output such as test scores,[16] with model estimation typically conducted using regression techniques (see for example Gustafsson, 2007; Van der Berg, 2008; Chetty & Moloi, 2011; Van der Berg, et al., 2011; Taylor, 2011; Spaull, 2013). The primary focus of much of the existing research has been descriptive and/or explanatory in nature; that is, determining important associations between the dependent and independent variables.[17] With descriptive and explanatory modelling, identification and estimation requires the researcher to make conjectures regarding the underlying relationships between inputs and the output of interest. If causal inference is not of primary concern, reliance on an underlying causal theory may be incorporated in a less formal way (Shmueli, 2010). However, one would still want to conduct a multivariate analysis that incorporates a combination of antecedent, mediator and moderator variables so as to at least arrive at a model that provides a close approximation to the true generative process. However, if the "true" relationship is not contained in the model, for example a true quadratic relationship modelled linearly, then any over- and/or underestimation resulting in different parts of the covariate space will lead to errors in inference (Barry & Elith, 2006).

An important issue that arises in education production modelling is that of missing covariates. Data limitations common to large-scale national surveys frequently result in the collection of a

---

Delegates (HOD) and black African schools were administered by the Department of Education and Training (DET) and each of the homelands had a separate education department. Regarding the use of the terms "white" and "black", I quote (from Spaull, 2012, footnote 2 and Coetzee, 2014, footnote 3): "The use of race as a form of classification and nomenclature in South Africa is still widespread in the academic literature with the four largest race groups being black African, Indian, coloured (mixed-race) and white. This serves a functional (rather than normative) purpose and any other attempt to refer to these population groups would be cumbersome, impractical or inaccurate".

[16] In the South African literature "non-linearities" have been introduced through hierarchical modelling (see for example Gustafsson (2007) and van der Berg (2008b)) that allows for random intercepts and/or slope coefficients. However, as with least squares regression the base model assumes linearity in the model parameters.

[17] Whilst "proper" statistical methodologies for testing causality exist, for example randomised experiments, in practice association-based statistical models applied to observational data are most commonly used for explanatory analysis. In cases where student performance is tracked over multiple years, omitted variable bias may be corrected for through the use of value-added modelling. Value-added applications in the South African context include Carnoy et al (2008), Taylor (2011) and Coetzee (2014).

"sufficient" set of covariates. Furthermore, indirect (distal) variables that are easily quantifiable and, to varying degrees, correlated with causal (proximal) variables are typically collected, even though ease of collection does not necessarily guarantee that the covariate will be free of measurement error. Omitted variables can produce discontinuities or multimodalities in the response surface, especially if the omitted covariate is correlated with specific values and/or ranges of the observed covariates. All of this implies that the response surface that needs to be modelled with the available data is likely to be more complex than the simple surface/s implied by theory. However, most studies attempt to approximate the response surface parametrically through simple components. In addition to missing covariates, other well documented issues with linear regression based models include: the order in which the predictors are introduced; multicollinearity; variable selection; outlier detection and removal; and model overfitting.[18]

This paper puts forward an argument in favour of flexible machine learning approaches for modelling education production as they allow for more complex response surfaces that are frequently observed in distributional data. Furthermore, in terms of viable modelling alternatives, they may be the most "natural for economic applications" (Varian, 2014). These techniques are primarily concerned with finding a function that is able to achieve good out-of-sample predictions. Predictive modelling is almost absent in economics, especially as a tool for developing theory. In fact, researchers might even regard prediction as unscientific. As stated by Berk (2008) "In the social sciences, for example, one either did causal modelling econometric style, or largely gave up quantitative work". In addition, the supposition made by some studies that a good explanatory model inherently contains some predictive power may come at the cost of making incorrect scientific and practical conclusions (Shmueli, 2010).

This study avoids starting with a data model, but rather uses an algorithm to *learn* the relationship between test performance and its determinants. The statistical technique of boosted regression tree (BRT) modelling as described by Friedman (2001) provides a highly flexible multivariate nonparametric regression technique that allows for nonlinear relationships to be fitted between covariates and the dependent variable without having to specify any functional relationship/s. There is mounting evidence in favour of using boosted regression over traditional linear regression models and other non-linear regression based techniques such as Generalised Linear models (GLM) and Generalised Additive models (GAM) (Bauer & Kohavi, 1999; Elith et al., 2008; Elith et al., 2006; Friedman, Hastie & Tibshirani, 2000; Friedman, 2001; Schonlau, 2005). To my knowledge, there exists no published examples of boosting and regression tree analysis applied to schooling outcomes data.

[18] See Hanushek (1979) and Todd and Wolpin (2003) for detailed discussions of these issues.

Assuming a complex and unknown data-generating process, BRT modelling attempts to learn the outcome through observing measured inputs and outcomes and finding dominant patterns with a focus placed on the model's ability to predict well. Regression trees perform automatic variable subset selection, which is useful for modelling schooling outcomes where a large number of potential predictors exist, yet only a few of them may be of actual relevance to prediction. The hierarchical structure of a tree further implies that interactions between covariates are automatically modelled without them having to be specified first. The advantage of such an approach for education production modelling is self-evident given that educational inputs typically do not have isolated effects but rather operate jointly in determining student performance. In addition, allowing for an unrestricted conditional expectations function whilst controlling for a multitude of covariates makes BRT better placed to bypass omitted variable issues related to specifically linear non-parametric estimation techniques. BRT modelling is further robust to model over-fitting, able to deal with missing values on controls and is (to a degree) immune to multicollinearity.

This study employs a large nationally representative school survey data set, the National School Effectiveness Study (2007-2009) that includes an indicator of former department. This variable is largely absent from other nationally representative South African datasets. I am therefore able to separate students into those that attended former DET and Homeland schools from those that attended former HOA, HOD and HOR schools. The data further allows for a multitude of potential predictors as it is particularly rich in terms of information regarding school management and classroom processes. The analysis begins with BRT analysis of the Grade 4 literacy and numeracy test scores within historically disadvantaged schools as well as visual investigations of the associations between predictors and the fitted response. The robustness of the main results are compared to random sub-samples of the full dataset as well as boosted models estimated for the 2009 wave of the NSES survey and the 2007 SACMEQ[19] survey dataset. The predictive performance of boosting is further assessed against linear and non-linear methods as well as competing machine learning methods.

The remainder of this chapter is structured as follows: section 2.2 details some of the existing research of the performance of former black schools in South Africa; section 2.3 describes the methodology employed; section 2.4 presents the data; sections 2.5 and 2.6 present the main empirical results and robustness checks respectively; section 2.7 concludes.

[19] Southern African Consortium for Monitoring of Educational Quality

2.2 The Performance of Disadvantaged Schools in South Africa

The South African education system may be described as one in which schools differ considerably in their ability to convert educational inputs into educational outcomes. Evidence has hinted towards a "bimodal" distribution of student performance; that is, a different data generating process for historically white schools than for historically black schools (Fleisch, 2008; Spaull, 2013; Taylor, 2011; Shepherd, 2013; Van der Berg, 2008). Spaull (2013) puts forward a twofold explanation for the bimodal pattern of performance. First, the historically black school system inherited from apartheid has remained largely dysfunctional and limited in its capacity to produce student learning, while the opposite is true of historically advantaged schools. Less affluent South African schools face both real and perceived constraints that inhibit effectiveness; "where communities are poor, have few material resources, and do not speak the language of instruction in their homes, there are few options to supplement the quality of teaching and learning in their schools" (Christie, Butler & Potterton, 2007: 101).

Secondly, the student and teaching bodies of these two school sub-systems are vastly different. Despite a distinct movement toward racial integration in historically advantaged schools, socio-economic integration has not occurred at the same level (Taylor & Yu, 2009). Socio-economic class has replaced race as the major determining factor of the social character or culture of a school. The movement of students has arguably occurred in a fairly predictable way as displayed by a "flight" of more affluent black students out of historically black schools, with little if any movement in the opposite direction (Chisholm, 2004).[20] Black schools are consequently left with the poorer members of the community. This may have effects on the educational performance of historically black schools, as the disadvantages faced by those from less affluent backgrounds are perpetuated through peer effects.

Kamper (2008: 2) argues that in order for historically disadvantaged schools to meet the challenges they face, some of which the education system was never designed to handle, they need to be innovative and creative in their schooling approach. Leadership styles such as being "visionary", perseverance, relentlessness, courage and risk-aversion have appeared as key factors for success (see for example Christie et al., 2007). Whilst it is encouraging to know that there are individuals within the public school sphere who possess these characteristics, these qualities are not

[20] An example of this is provided in an article by Woolman and Fleisch (2006). They describe how Sandown High in Sandton (a relatively high income urban suburb of Johannesburg) is oversubscribed, whereas on the other side of town in Orlando High Soweto (a township on the outskirts of Johannesburg) classrooms stand empty. Many of the students attending Sandown High reside close to Soweto, yet they choose to travel many kilometres to attend school elsewhere.

easily replicated. Taylor (2008) puts forward two key issues which, if addressed, may lead to improved outcomes in former disadvantaged schools, namely time management and curriculum leadership.

In terms of time management, principals have been quick to blame forces outside of their control (e.g. public transport) as contributors to high levels of teacher and (to a lesser degree) student absenteeism, indicating an underlying failure on the part of school management to "take responsibility and exercise control over their own work environment" (Taylor, 2008: 7). In two separate qualitative studies of historically disadvantaged schools who performed well in the school leaving examinations, Malcolm et al (2000) and Christie et al (2007) find that time was of highest priority as displayed by strict punctuality (sticking to the timetable) and extended school hours. Time management in terms of ensuring that teachers are devoting the required number of hours to teaching is of further importance. Utilising the NSES panel data, Taylor (2011) finds substantial gains in student learning when teacher knowledge is combined with time-on-task. Hallinger and Murphy (1986) find that effective poor schools are more likely to maximize the amount of time allocated to basic skills instruction during school time and make less use of homework. Teaching processes such as these may compensate for the lack of school preparedness of students, as well as a lack of time available for independent study outside of school.

International research has revealed that a student's own motivation to read outside of school is important to the process of becoming literate (Chapman & Tunmer, 2003; Linnakyla, Malin & Taube, 2004). The Pupil Progress Project (PPP) study undertaken in the Western Cape province of South Africa in 2003 indicated that children who frequently read and engaged with homework outside of school hours performed significantly better on reading and literacy tests. Using the PIRLS (2006) dataset, Shepherd (2011) finds a significant and positive association between reading scores and the frequency of and time spent on reading homework for the sample of African language testing schools (as a proxy for the former black African school department).

Poor schools may also rely more heavily on providing students with tangible (extrinsic) rewards for their classroom accomplishments in order to instil motivation and confidence (Hallinger & Murphy, 1986: 345). In a qualitative assessment of two above average performing disadvantaged schools in the Western Cape province, Wilburn (2013) finds that teaching and learning are regulated through forms of high expectations. In the one case, a learning culture is fostered through a broader social expectation of quality education from the community such as that the students might one day contribute positively to society. This expectation is supported by an ideology that, with the appropriate support, all students are capable of achieving. The creation of a school community that breeds a sense of acceptance and worth can help students accept and commit to shared educational

values (Dauber & Epstein, 1993). We could expect this to be more pronounced for less affluent students who may experience a lack of similar support at home. The establishment of such a community not only conveys a broader set of values that are concerned with mutual respect and appreciation, but also motivates individuals within the community to abide by these values (Battistich, Solomon & Kim, 1995).

2.3 Empirical approach

2.3.1 Regression trees

With its roots in computer science, regression trees have become a popular data mining technique used in statistics and machine learning (Friedman et al., 2000; Friedman, 2001; Morgan & Sonquist, 1963; Ridgeway, 1999). This paper will discuss modern regression trees as described by Breiman, Friedman, Olshen and Stone (1984). Consider a sample of *i* = 1,…, N observations with known values $\{y_i, \boldsymbol{x}_i\}$ where *y* is a random output variable and $\boldsymbol{x} = \{x_1, \dots, x_n\}$ is a set of random predictor variables which may be of any type (numeric, categorical, binary and ordinal). The measurement space χ is taken to be the set of all possible predictor values and let $C = \{c_1, \dots, c_J\}$ be the set of possible classes. A classifier (such as a regression tree) can then be defined as a function $F(\boldsymbol{x})$ with domain χ and range C that corresponds to a partition of χ into J disjoint regions where a constant, such as the sample average outcome, is fit to the elements of that region. That is:

$$\boldsymbol{x} \epsilon R_j \Rightarrow T(\boldsymbol{x}; \{R_j\}_1^J) = \bar{y}_J \qquad [2.1]$$

where $T(\boldsymbol{x}; \{R_j\}_1^J)$ represents a regression tree model comprised of J disjoint regions and $\bar{y}_J = \frac{1}{|R_j|} \sum_{\boldsymbol{x}_i \epsilon R_j} y_i$ are the values below each terminal node (model coefficients). In general, we wish to obtain an estimate $\hat{F}(\boldsymbol{x})$ of $F(\boldsymbol{x})$ such that the expected value of some specified loss function $L(y, F(\boldsymbol{x}))$ is minimised:

$$\hat{F}(\boldsymbol{x}) \cong F(\boldsymbol{x}) = \arg\min_{F(\boldsymbol{x})} E_{y\boldsymbol{x}}\, L(y, F(\boldsymbol{x})) \qquad [2.2]$$

The regression tree is constructed through making repetitive splits of χ so that a hierarchical structure is formed. The complexity of the regression tree is determined by the number of splits, where each split allows for additional interactions between variables. It should be noted that when dividing χ into subsets, any subsequent partitions on these subsets do not have be performed on the same variable, nor does the tree have to be symmetric. This allows for a heterogeneous response model.

The general goal in dividing χ is to make the distributions of elements across classes different in such a way that, with respect to *y*, the data corresponding to each child node is purer than the data corresponding to the parent node. The algorithm executes a comprehensive search through all predictors as well as all values of the predictors in order to maximally reduce variability in the response. This can yield a bias in variable selection as the so-called greedy algorithm tends to choose categorical variables that have many distinct values as a splitter (Loh, 2002; Qin & Han, 2008). An ordered (continuous) predictor x_1 with n distinct values can give rise to (n-1) potential binary splits of the data. If we consider two ordered predictors, x_1 and x_2, with n_1 and n_2 distinct values respectively, and $n_1 > n_2$, then all else constant, $x_1$will have a higher chance to be selected than x_2. A selection bias towards predictors that take on many values can lead to erroneous inferences being drawn from the tree structure as some other split on another variable may have led to more effective further splitting; that is to say, locally optimal decisions do not guarantee a globally optimal decision tree.[21] Multicollinearity is a further issue for variable selection as when two variables both explain the same thing, a decision tree will greedily choose the best one. Ensemble methods such as boosting, discussed next, can negate this to a certain extent, although at the cost of ease of interpretation.

2.3.2 Boosting and regularisation

Boosting, coupled with regularisation methods, is able to mitigate the issues of regression tree analysis as well as improve model accuracy. Boosting is a method that adds together many simple functions to estimate a smooth function of a large number of covariates (Schapire, 2003). In the context of this study, each simple function is a regression tree. Boosted regression for a continuous, normally distributed outcome variable can be described by a gradient boosting algorithm that aims to minimise a loss function at each step (iteration) by adding a new tree that best reduces the loss function (Friedman, 2001). This study makes use of the boosted regression tree algorithm as laid out in Friedman and Meulman (2003), of which a simple summary is provided by Schonlau (2005: 336).

The first regression tree $F_0(\boldsymbol{x})$ is grown on the sample $\{y_i; \boldsymbol{x}_i\}$ such that the residuals are minimised. Subsequent iterations use the residuals left over from the previous iteration as the response variable; that is, for the proceeding *m* = 1,…,M iterations, the BRT model consisting of all previous regression trees is updated to reflect the current regression tree, and at each step the residuals are updated to reflect changes in the BRT model. For the first iteration, we grow the

[21] Conditional inference (CI) trees is one method by which biased variable selection can be avoided in constructing regression trees (see for example T Hothorn, Hornik and Zeileis, 2006). However, no statistical package currently exists which combines CI trees and boosting, although CI trees are combined with random forests in the "party" package in R (Hothorn, Hornik, Strobl and Zeileis, 2014).

tree F_1 using the residuals $r_{1i} = y_i - F_0(\boldsymbol{x})$ and the covariates $\boldsymbol{x}$. The regression tree F_1 is then added to the current best fit $F_0(\boldsymbol{x})$ to re-estimate the fitted outcome for each observation $F_1(\boldsymbol{x})$. This is known as a *forward stagewise* fitting procedure. For the second iteration we grow a tree using the residuals $r_{2i} = y_i - F_0(\boldsymbol{x}) - F_1(\boldsymbol{x})$, which is then added to $F_1(\boldsymbol{x})$, and so on. The final model is therefore a linear combination of many trees. This process repeats until a stopping criterion is reached. Subsequent trees in the algorithm process are not restricted to contain the same predictors as previous trees nor do the split points on predictor variables have to be the same. However, the size of the trees F_m grown at each iteration is fixed ahead of time.

In fitting the BRT model, two parameters need to be specified. First, the number of splits that will be used for each regression tree (the number of interactions). This is also referred to as the tree complexity, tc. Specifying *J* splits corresponds to a model with up to *J* -way interactions as *J* covariates need to be considered jointly. The second parameter is the learning rate (shrinkage) parameter, lr, which reduces the impact of each additional tree. Shrinkage is accomplished by introducing a parameter λ as follows (Schonlau, 2005):

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \lambda * (\text{last regression tree of residuals}) \qquad [2.3]$$

where $0 < \lambda \leq 1$. Stochasticity is introduced into the model through "bagging" which specifies that only a random subset of the residuals is selected to build the regression tree at each iteration. This is thought to reduce the variation of the final prediction without affecting bias as all residuals will be used across all trees (Friedman, 2001). Elith et al (2008) show that bagging improves model accuracy and reduces overfitting. For purposes of this study, we use a bag fraction of 0.5.

Regularisation methods are used in order to strike the best balance between model fit and predictive performance (Hastie, Tibshirani & Friedman, 2009). Essentially, this involves jointly optimising *lr*, *tc* and *nt*. Too many iterations will result in over-fitting; too few iterations will lead to a poorly fitted model. A smaller *lr* implies a larger number of iterations. In general, a smaller *lr* and a larger *nt* are preferable, dependent on the sample size. The *tc* also affects the optimal *nt*, as the more complex the underlying tree, the lower the *lr* required for optimising the loss function. Therefore, increasing the model complexity requires decreasing *lr* (usually inversely) in order to keep *nt* unchanged. In theory, the tree complexity should reflect the true interaction order in the response being modelled (Elith et al, 2008). However, there are gains to increasing *tc* when the sample size is large. In the case of small samples, however, the outcome is best modelled using simple trees and a slow enough *lr* so as to allow for at least 1000 iterations, the recommended minimum for fitting BRT models (Elith et al., 2008).

One approach for selecting optimal model settings is cross-validation (CV). CV provides a means of testing the model on withheld portions of the data while still using all data to fit the model. This is similar to using a portion of the data to fit the model (training data) and the remaining data for model prediction (test data). In a five-fold CV, for example, the data set is split into five discrete subsets of 20% of the data. Each subset is then used as test data and the remainder as training data. In order to determine the predictive accuracy of the model, a pseudo R^2 is computed on both the training and test data sets where:

$$R_{DEV}{}^2 = 1 - \frac{\text{mean residual deviance}}{\text{mean total deviance}} \quad [2.4]$$

Similar to the familiar R^2, the pseudo R^2 is interpreted as the "fraction of variance explained by the model".

2.3.3 Interpretation

As BRT models are based on a linear combination of many trees, the results are not easily interpretable. With BRT analysis we focus on the relative importance, or influence, of individual predictors in predicting the outcome of interest using formulae developed by Friedman (2001). The measures are based on the frequency with which a predictor is selected for splitting, weighted by a squared improvement to the model as a result of each split, and averaged over all iterations (Friedman & Meulman, 2003). This can be expressed as:

$$I_j^2 = \frac{1}{M}\sum_{m=1}^{M} I_j^2(T_m) \quad [2.5]$$

where I_j represents the relevance of predictor x_j. The influence of each predictor variable is standardised so that the sum adds up to 100 percent. As a regression tree is not able to separate main and interaction effects, the influences defined in equation [2.5] are not able to say anything about the direction or magnitude of the relationship of the variable with the outcome. This is unlike a linear regression approach typically used for modelling education production.

However, we are able to visualise the effect of a predictor through partial dependence plots. While these may not be perfect representations of the effects of each predictor - particularly if the underlying function is dominated by higher-order interactions and strong correlations – they provide a useful basis for interpretation. The partial dependence of a predictor x_k can be estimated by:

$$\hat{F}_k(x_k) = \frac{1}{N}\sum_{i=1}^{N} F(x_k, \boldsymbol{x}_{-k(i)}) \quad [2.6]$$

where $\boldsymbol{x}_{-k(i)}$ denotes the data values of all other predictors. $\hat{F}_k(x_k)$ is then the effect of x_k on the outcome holding all other variables at their average. Here again we see a difference with the

regression interpretation of partial regression coefficients where the effect of all other covariates, $x_{-k(i)}$, are ignored. Only in the unlikely event that x_k and $x_{-k(i)}$ are independent will the partial dependence as described by [2.6] be equivalent to the marginal effect.

In a similar fashion we can quantify the nature and size of interactions between two predictors. The H statistic (Friedman & Popescu, 2008) provides a measure of interaction strength. Essentially, if two variables x_k and x_j do not interact with each other, then $F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k)$ (Lampa, Lind, Lind & Bornefalk-Hermansson, 2014). The statistic H_{jk} captures the proportion of variance of $F_{jk}(x_j, x_k)$ that is not captured by $F_j(x_j) + F_k(x_k)$. H_{jk} ranges from 0 to 1, with larger values signalling stronger interactions. It should be cautioned that sampling fluctuations can lead to spurious interactions; therefore one should be aware that a non-zero value of H may not reflect a true interaction. Unfortunately there exist no formal rules for assessing interaction significance in the context of BRT modelling so distinguishing between low and higher order interactions is not possible.

All fitted BRT models and graphing for this analysis are obtained using the "gbm" (Ridgeway, 2007) and "dismo" (Hijmans, Phillips, Leathwick & Elith, 2011) libraries in R. Model parameters are selected using the "caret" library in R (Kuhn, 2008).

2.4. Data

Data for the National School Effectives Study (NSES)[22] was collected between 2007 and 2009 on a nationally representative sample of schools in South Africa.[23] Unlike most school survey data collected in South Africa, the NSES provides an indicator of former school department and school poverty quintile[24] for each school. We are therefore able to easily separate schools into historically disadvantaged (former DET and H) schools and historically advantaged (HOA, HOD and HOR) schools. Students in 266 schools were tested in literacy and numeracy in 2007 (Grade 3), 2008 (Grade 4) and 2009 (Grade 5).[25] This paper focuses primarily on the 2008 Grade 4 sample. As a universal sample of students was taken from the respective grades in each year, the sample sizes are large at approximately 16000 students per year. The same tests were administered at all grades making the results comparable from one year to the next.

[22] Managed by JET Education Services and funded by the Royal Netherlands Embassy.
[23] Schools from the Gauteng province were not surveyed as the province was engaging in their own independent test at the same time. School numbers by province were randomly sampled such that the distribution mirrored that found within the national school list of ordinary public schools. Once sampled, all Grade 4 students in all Grade 4 classes were surveyed.
[24] All Public Ordinary Schools in South Africa are classed into one of five quintiles. These are determined by analysing socio-economic indicators of the communities surrounding the school. As of 2012, the poverty quintile classification will be replaced by the classification of schools as either fee paying, or non-fee paying.
[25] The same students were tested in each year thus producing a panel dataset. However, due to attrition, only 8383 students were captured in all three waves, approximately 55 percent of the annual samples.

In addition to student testing, a wide variety of contextual information was collected through student questionnaires, teacher questionnaires and school principal questionnaires. The coverage of issues relating to school and classroom processes was remarkably detailed for a sample survey of this size. For example, an extensive document review was carried out including an examination of the frequency with which various types of exercises appeared in student workbooks. English teachers were further asked to take a short literacy test and mathematics teachers took a short numeracy test. Although this may only be a crude measure of teacher subject knowledge, it may provide a proxy for teacher quality. Derived from the contextual questionnaires, the control variables in the BRT models for numeracy and literacy are a mixture of continuous, ordered and binary; brief descriptions of these are provided in table A1 of the appendix.

Accounting for missing data on student age and gender, the sample consists of 14408 grade 4 students in 251 schools. Observations are split on former department classification as follows: 11894 students in 209 former DET and Homeland (H) schools and 2514 students in 42 former HOA (white), HOD (Indian) and HOR (coloured) schools. This division is in line with the South African school population. The distributions of numeracy and literacy scores for DET/H and HOA/HOR/HOD schools in 2008 are displayed in figure 2.1. The maximum scores on the numeracy and literacy tests were 51 and 58 points, respectively[26]. Average test scores in HOA/HOR/HOD schools are approximately 1 to 1.5 standard deviations higher than in DET/H schools. Filmer et al. (2006) compare a years' worth of learning to approximately 0.4 to 0.5 standard deviations on a standardised test. A difference of 1.5 standard deviations would therefore be equivalent to 3 to 4 years of learning which appears quite large in the context of a Grade 4 test. Spaull and Kotze (2014) find that the learning gap between the poorest 60 percent and wealthiest 20 percent of South African Grade 3 students is approximately 3 grades. A wider spread of test scores amongst HOA/HOD/HOR schools is evident, primarily due to the relatively weaker performance of HOR schools that are on average poorer and less resourced than HOA and HOD schools.[27]

2.5 Results

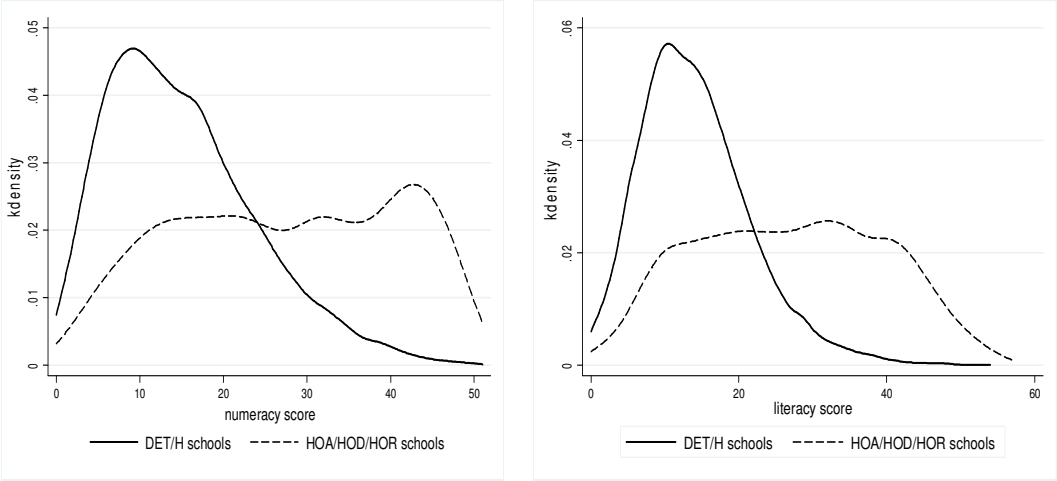
2.5.1 From single to multiple tree regressions

As an illustration of the underlying process of a boosted regression tree model, figure 2.2 presents the tree structures fitted at the beginning stages of the iteration process. Taking the default model parameters as discussed in section 2.3, I begin by fitting a BRT model with *nt* = 1000, *lr* = 0.05 and *tc*

[26] We could have similarly used the percentage score on the tests as the dependent variable. However, this would not change the interpretation of the results.
[27] More than half of HOR schools are classified within the bottom three school poverty quintiles.

= 4. A 50 percent bag fraction is also adopted. The first two trees shown in panel (a) of figure 2.2 have two of four variables in common with splits occurring at slightly different values.[28] As each tree is allowed to differ, one can see how a final model comprising of many trees allows for a heterogeneous response function. Panel (b) of figure 2.2 illustrates how the boosting process fits a non-linear response. A partial dependence plot of school SES from the first tree split shows up as a small step; adding information from the second tree adds a second step. As more trees are included in the partial plot, the response to school SES becomes more complex and curvilinear (Elith et al, 2008).

Figure 2.1: Kernel densities of grade 4 numeracy scores in 2008, by former school department



Notes: own calculations using NSES 2008; DET = Department of Education and Training, H = homeland schools, HOA = House of Assembly (white) schools, HOD = House of Delegates (Indian) schools, HOR = House of Representatives (coloured) schools.

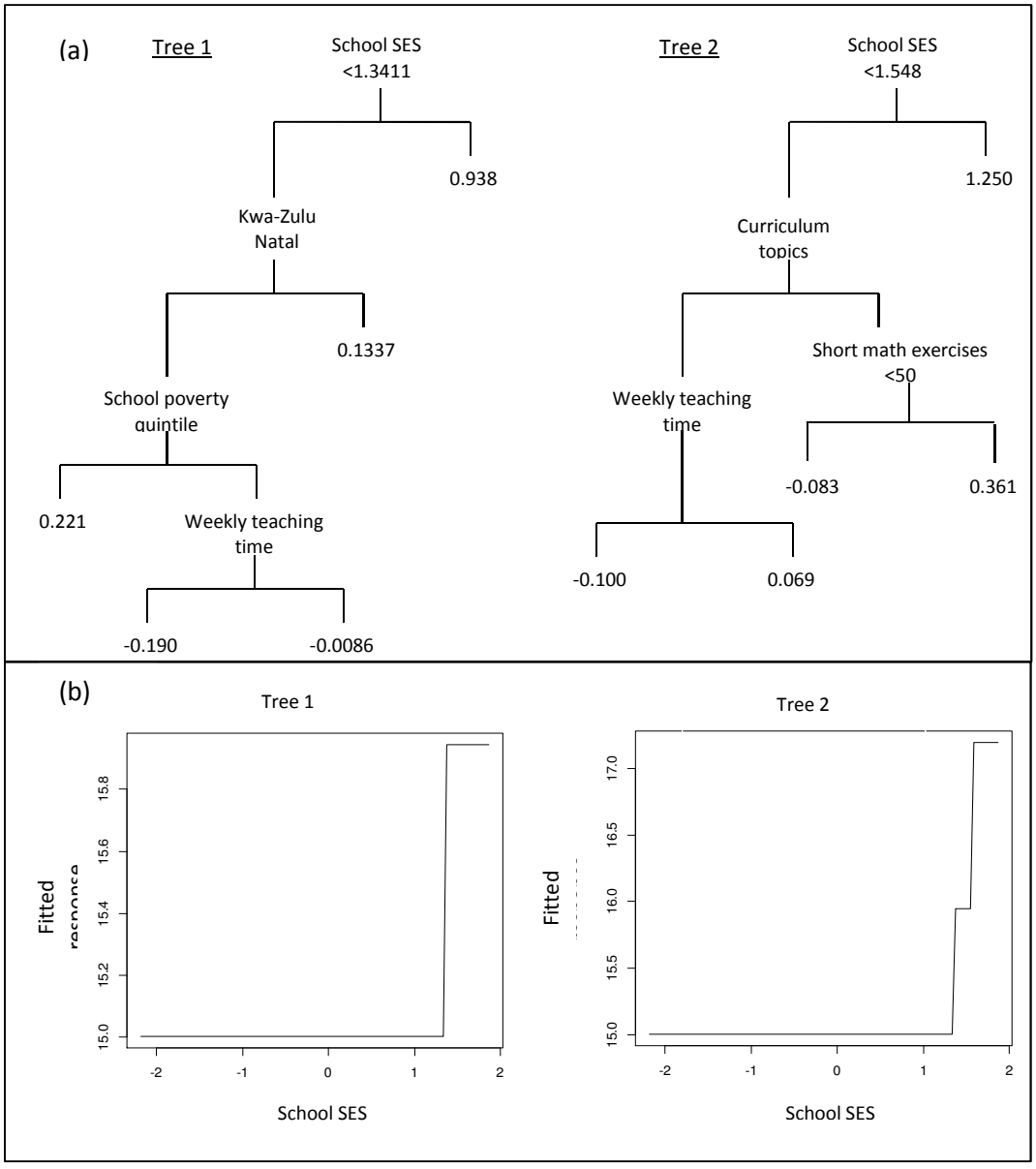
2.5.2 Tuning of model parameters

I begin the analysis by first determining the combination of *tc*, *lr* and *nt* that achieves a minimum out of sample predictive error. Figure A2.1 of the appendix to this chapter shows the predictive deviance (represented by the root mean squared error (RMSE)) against *nt* for varying tree complexities holding *lr* constant at 0.10. Higher degrees of *tc* are found to be related to fewer *nt* in order for minimum error to be reached. A model with *tc* = 1 (decision stump) is observed to perform the worst. Improvements in predictive accuracy with more complex trees is expected in larger samples as the complexity of information contained within multiple observations can be better modelled by more complex trees (Elith et al, 2009). Table 2.1 summarises the predictive accuracy,

[28] Note that the learning rate has indeed resulted in predicted outcomes represented at the terminal nodes that are small relative to the size of the final test score.

determined by cross-validation estimates of RMSE and R-squared, of BRT models based on different combinations of tc and lr. Although models with *tc* = 4 have a better in- sample performance, they require low *lr* and many *nt* in order for the minimum model error to be reached. The loss in performance when estimating models with less complex trees is not dramatic, with no notable difference observed across models with tree complexities of 2 and 3. In fact, the out-of-sample performance is the same regardless of the size of *tc*. Given these findings, a computationally less demanding strategy is adopted with *tc* = 2 and the fastest (largest) *lr* that achieves $nt \geq 1000$ is selected. With regards to Grade 4 numeracy, the BRT model uses a learning rate of 0.10 and 1950 iterations, whilst the Grade 4 literacy model is built using a learning rate of 0.10 and 1450 iterations.

Figure 2.2: Example tree structures and partial plots from a BRT model



Notes: own calculations using NSES 2008 and the gbm package in R.

Table 2.1: Predictive performance across model parameters

Numeracy score model						
Tree complexity	4	4	3	3	2	2
Learning rate	0.025	0.05	0.025	0.05	0.05	0.10
Number of iterations	2750	1950	3050	1550	2750	1950
RMSE	6.38	6.40	6.45	6.48	6.59	6.49
R-squared	0.48	0.48	0.47	0.47	0.45	0.46
CV RMSE	6.90	6.92	6.91	6.91	6.93	6.93
CV R-squared	0.39	0.39	0.39	0.39	0.39	0.39
Literacy score model						
Tree complexity	4	4	3	3	2	2
Learning rate	0.025	0.05	0.025	0.05	0.05	0.10
Number of iterations	2900	1450	3350	1900	2900	1450
RMSE	5.41	5.42	5.48	5.44	5.57	5.58
R-squared	0.47	0.47	0.45	0.46	0.43	0.43
CV RMSE	5.89	5.89	5.88	5.88	5.90	5.91
CV R-squared	0.36	0.37	0.37	0.37	0.36	0.36

Notes: own calculations using NSES 2008. Cross-validation (CV) is performed using 10 folds. All models use a 50 percent bag fraction.

2.5.3 Relative influence of control variables and partial dependence plots

The relative influence of the 15 most important model predictors across the Grade 4 numeracy and literacy models are summarised in table 2.2.[29] As mentioned in section 2.3, the relative influence of each predictor is scaled so that the total equals 100. Therefore, higher percentages reflect greater importance in the model. Recall that this in no way reflects the magnitude or direction of the relationship between the predictor and the response of interest. It is worth noting that 9 of the 15 most influential predictors are the same across the two models, with some differences in relative ranking; the combined importance of these common predictors is 41.85 percent for numeracy and 48.13 for literacy. Where differences occur, these are largely subject specific but appear to be indicative of the same underlying factor; for example, the frequency of exercises specific to mathematics or reading. School SES comes through as the most important predictor of performance in both tests. This is fairly unsurprising given that school SES is generally thought to be a catchall variable of overall school quality and access to resources. School and classroom factors appear to be relatively more "influential" than home background in predicting Grade 4 performance.

[29] Observation weights could be generated for the NSES sample that weight up the student numbers to be representative of those found within each province; that is, a different student weight by province, but the same student weight within province. The gbm package in R allows for the inclusion of "site weights" in the BRT model. The analysis of this paper using the NSES was also estimated using these student weights with no obvious difference in results. Therefore, the results shown are unweighted.

Table 2.2: Summary of the relative contributions (%) of controls for boosted regression tree models of numeracy and literacy test scores

Grade 4 numeracy model		Grade 4 literacy model	
School SES	8.08	School SES	11.09
Curriculum topics covered	7.87	School pupil-teacher ratio	6.76
Short math exercises	7.23	Household SES	5.91
Class size	6.29	Class size	5.75
School pupil-teacher ratio	5.75	Teacher experience	5.73
Teacher's weekly teaching time	5.15	Sentence writing more than ½ page long	5.60
Household SES	4.68	Age	4.70
Intermediate Phase math classes (weekly hours)	3.68	Frequency watch television in English	3.80
Teacher experience	3.53	Teacher's weekly teaching time	3.66
Age	3.37	Word exercises less than ½ page long	3.46
Complex math exercises	3.12	Paragraph exercises less than ½ page long	2.65
Long math exercises	3.03	Female	2.42
Frequency of reading homework	2.69	Word exercises more than ½ page long	2.34
Kwa-Zulu Natal	2.63	Frequency read alone at home	2.27
Frequency read alone at home	2.31	Frequency of reading homework	2.26
Number of iterations	1950	Number of iterations	1450
Shrinkage	0.10	Shrinkage	0.10
Tree complexity	2	Tree complexity	2
RMSE	6.49	RMSE	5.58
R-squared	0.46	R-squared	0.43
Observations	11894	Observations	11894

Notes: own calculations using NSES 2008. Models are developed using 10 fold cross-validation and 50 percent bagging.

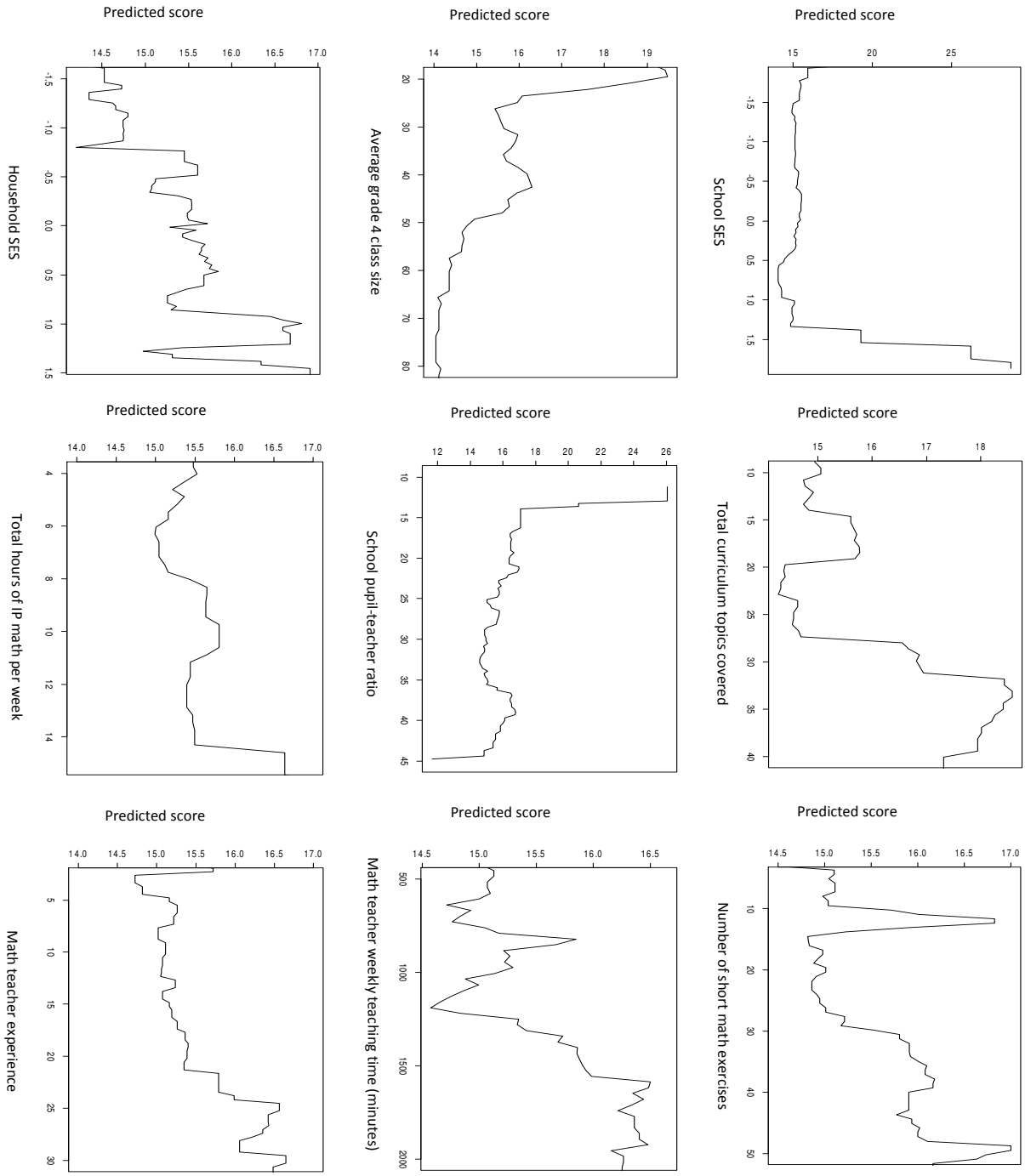
Visualisations of the relationships between the most influential predictors and estimated test scores are achieved through partial dependence plots that illustrate the effect of a chosen predictor on the fitted outcome holding all other variables at their average. These are indicated in figure 2.3 and figure 2.4. Although the step appearance of the plots may not lend itself to a natural interpretation, there is evidence of non-linearities in the relationships between predictors and student achievement it should be kept in mind that any strong interactions and/or correlations in the data may influence the shape of the plots, including correlations with omitted variables. School SES is observed to have a similar relationship with test scores across both subjects; that is, fairly flat until approximately 1 to 1.5 standard deviations above average after which the positive slope steepens dramatically. As the shape of the partial dependence plot may be influenced by interactions of school SES with other model covariates, the higher expected performance of students attending wealthier schools may be related to simultaneous access to complimentary schooling inputs. It is understood that schools with higher concentrations of low SES students are more likely

to suffer from infrastructional and resource shortages, particularly access to high quality teachers and small class sizes. The non-linearity in the relationship between school SES and performance at higher values of school wealth may further be indicative of a correlation between school wealth and omitted school quality variables over this range of school SES.

The partial dependence plots also provide evidence that opportunity to learn (OTL) and time-on-task (TOT) are fundamental for creating augmented performance in former black schools. Higher frequencies of classroom exercises are related to better performance, as is coverage of a greater portion of the core curriculum. The spike observed at approximately 10 counts of short math exercises[30] should not be taken to suggest that fewer of these types of exercises is better, but rather that a teacher who places less emphasis on these types of exercises is possibly engaging students with more complex calculations, and vice versa. It is interesting to note that positive returns to class exercises only appear once a high (above average) threshold is reached. Contact time with teachers is further observed to be positively related to performance, more so in the case of mathematics. Teachers are expected to have formal contact teaching time in the region of 25 to 35 hours per week (Department of Education, 2002). The highest predicted math scores are estimated for students taught by teachers who report formal in-school teaching hours within this band.[31] The negative relationship between class size and pupil-teacher ratios further suggests that overcoming the binding constraints of overcrowding and lack of teachers is associated with better outcomes. Teacher experience is also observed to be positively related to test scores, notably so after 20 years of teaching experience. At the household level, greater exposure to English (the test language) through the medium of television is related to augmented literacy test results. This effect may be two-fold as daily exposure to the test language is likely to increase familiarity with the subject content, but also the availability of television may be indicative of the affluence of the home environment. Household SES is also evidenced to be positively and (approximately) linearly related to test scores.

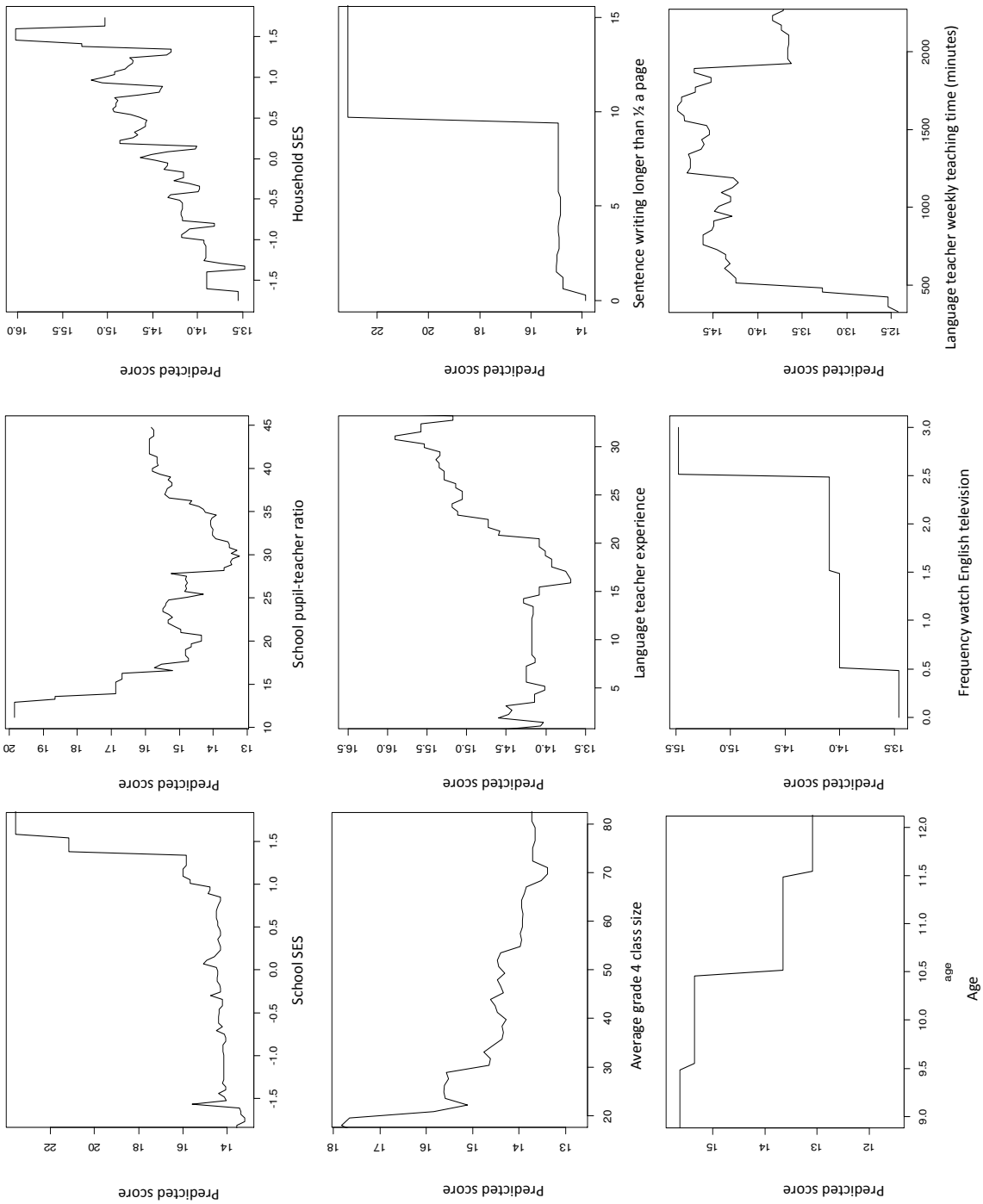
[30] Short math exercises are defined as being 5 lines or less.
[31] In the case of language teachers the results are less definitive in that the plot suggests positive gains for students taught by language teachers reporting to teach for at least 8 or more hours a week.

Figure 2.3: Partial dependence plots for the nine most influential variables in the model for grade 4 numeracy



Notes: own calculations using NSES 2008. Fitted responses (y-axes) are estimated assuming average values on all other controls except that plotted on the respective x-axis. For explanation of variables and their units see table A2.1 of the appendix to this chapter.

Figure 2.4: Partial dependence plots for the nine most influential variables in the model for grade 4 literacy



Notes: own calculations using NSES 2008. Fitted responses (y-axes) are estimated assuming average values on all other controls except that plotted on the respective x-axis. For explanation of variables and their units see table A2.1 of the appendix to this chapter.

The relationship between performance and SES (school and household levels) identified by the analysis of this paper is dissimilar to those reported in previous research. Taylor and Yu (2009) investigate the relationship between SES and test outcomes using the PIRLS 2006 dataset and information on the language of testing (chosen by the school according to the foundation phase LoLT) and home language of students to identify a crude proxy for former school department.[32]

Locally weighted smoothing applied to the data for the two separate sub-systems indicated a SES gradient that was relatively flat at all levels of school SES for the group of African language testing schools. The difference in results between this paper and that of Taylor and Yu (2009) may be related to the proxy for former department. Figure A2.2 of the appendix shows test performance plotted against school SES using kernel-weighted local polynomial smoothing for three separate groups of schools: (i) former HOA/HOD/HOR schools, (ii) former DET and Homeland schools that reported their foundation phase LoLT as English and/or Afrikaans and (iii) former DET and Homeland schools that reported an African language as the foundation phase LoLT. It is immediately clear that the high average SES former DET and Homeland schools who teach in either English and/or Afrikaans from Grade R drive the nonlinear relationship identified in this paper. A similar result is found for numeracy performance against household SES (results not shown here). As mentioned the non-linearity may be reflective of omitted school quality factors. In an assessment of the effect of language of instruction on performance, Taylor and Coetzee (2013) find that failure to correct for confounding factors such as omitted school quality factors leads to an upward bias in the relationship between English instruction in the foundation phase grades and reading performance in grades 4, 5 and 6. However, controlling for school fixed effects results in the converse result; that is, a significant improvement in performance linked to mother-tongue instruction in the first four years of school.

2.5 Identification of important interactions

The ten two-way interactions with the highest H_{jk} statistic from each of the Grade 4 models are reported in table 2.3.[33] As interest is primarily in the interaction between school level processes and/or resources, interactions between home background factors will not be investigated in great detail.[34] Table A2.2 of the appendix summarises the strength of the relationship between the

[32] Schools that tested in an African language were classified as former disadvantaged, while schools that tested in English or Afrikaans with at least 25 percent of students reporting speaking the test language at home were classified as former advantaged schools. Schools that tested in English or Afrikaans but had more than 75 percent of students with a home language other than English or Afrikaans were excluded from the analysis.

[33] The numeracy and literacy models yield 387 unique two-way interactions each, although the majority of interactions have a H_{jk} statistic that is smaller than 0.02.

[34] It is, however, interesting to note that the strongest interaction in both models occurs between the two factors of adult reading. This result may indicate a spurious interaction, which would be unsurprising given that a zero

variables found to interact in the BRT models. In the case of continuous variables a Pearson's correlation coefficient is reported, whilst a Pearson's chi-squared test is used in the case of two categorical variables. The two indicators of adult reading behaviour are found to be strongly correlated. In instances such as this, Friedman and Popescu (2008) advise that one should avoid entering such spurious interactions into the predictive model, or at least avoid reporting them. Most of the strongest two-way interactions identified in the Grade 4 literacy model are found to exist between two variables that are highly correlated. Furthermore, indicators of household and regional characteristics that correlate to each other as well as to language outcomes come through as strong two-way interactions, which may be related to the fact that the language test was written in English only. Ten percent of former homeland and DET schools sampled in the NSES reported either English and/or Afrikaans as the language of teaching and learning (LoLT) in the Foundation Phase, with all schools switching to English and/or Afrikaans in Grade 4.

Figures 2.5a–2.5f illustrate the joint partial dependence (contour) plots of variables found to interact in the numeracy model.[35] The number of short math exercises found in student workbooks is evidenced to positively interact with the number of curriculum topics covered (figure 2.5a). A greater number of each of these variables in their own right is positively related to performance, but in combination is related to the highest predicted test result, all else equal. A similar finding holds for the interaction between curriculum coverage and teacher experience (figure 2.5c) as well as math teacher test score and teacher experience (figure 2.5b). It is noteworthy that regardless of teacher experience, a greater coverage of curriculum is positively related to performance. Similarly, regardless of curriculum coverage, a student taught by a very experienced teacher is predicted to perform better. This result is encouraging in a schooling context where teacher experience might be lacking. However, one might argue that coverage of the curriculum is dependent on experience. Further investigation indicates that whilst the relationship is indeed positive, it is weak. The results shown here suggest that training targeted at providing teachers with the pedagogical skill necessary to identify which aspects of the national curriculum require the most attention at different phases of primary school learning may improve the performance of students in former disadvantaged schools. In a recent study of teacher knowledge of the mathematics curriculum over Grades 3-9 in the Gauteng province, Shalem, Sapire and Huntley (2013) found discrepancies between what teachers understood as intended by the national curriculum and what they reported having enacted in their

response on the one factor ("an adult doesn't read to me at home") almost perfectly predicts a zero response on the other ("an adult never reads to me").

[35] Graphical visualisation of the fitted function for two interacting variables is one advantage that the "dismo" and "gbm" libraries have over competing ensemble method techniques such as random forests and Bayesian additive regression tree models.

classrooms, particularly in grades 3, 4, 5 and 6 where 44 percent of the content classified by teachers as “not taught” was at the expected grade level.

Table 2.3: Strongest two-way interactions for Grade 4 numeracy and literacy BRT models

Numeracy model		
Variable 1	Variable 2	H-statistic
Frequency adult reads to student	Adult reads at home	0.7114
Short math exercises	Curriculum topics covered	0.3324
Teacher experience	Math teacher test score	0.3290
3 or more children in the home	Western Cape province	0.3252
Curriculum topics covered	Teacher experience	0.2500
Frequency adult reads to student	Western Cape province	0.2085
Weekly teaching time	Household SES	0.2066
LOLT textbooks for all students	Curriculum topics covered	0.1916
Curriculum topics covered	Hours of IP math per week	0.1861
Weekly teaching time	Frequency student reads at home	0.1822
Literacy model		
Variable 1	Variable 2	H-statistic
Frequency adult reads to student	Adult reads at home	0.8249
Staff computers present and functional	North West province	0.7226
School poverty quintile	Help with homework from father	0.6175
Zero teachers absent	Electricity present and functional	0.6156
Weekly teaching time	Teacher experience	0.6082
Home language African	3 or more children in the home	0.4290
Toilets present and functional	Only child in the home	0.3979
Shortage of LTSM	Permanent principal	0.3463
LTSM unused	Speak English regularly at home	0.3426
3 or more children in the home	North West province	0.3346

Notes: own calculations using NSES 2008.

Figures 2.5d-2.5f further illustrate the important role that TOT and OTL play in determining outcomes in former disadvantaged schools. There is a delicate interrelationship between TOT and OTL; students will not be able to exhibit learning if they have not been provided with enough OTL and teachers cannot be expected to complete the curriculum if there is not sufficient TOT (OECD, 2008: 173). Figure 2.5e supports this claim. The policy mandated allocation to mathematics instruction at the Intermediate Phase (IP) is approximately 3.5 hours a week per grade (Department of Education, 2002). Therefore, total IP math instruction should be 10 to 11 hours a week. Students taught in classrooms where less than a third (±28 topics) of the core curriculum is completed achieve lower results, all else equal, unless the school reports at least 5 hours of mathematics instruction for grade 4 students (15 hours in total across the three IP grades). A combination of 5 hours of grade 4 math instruction coupled with a, relatively speaking, high coverage of the curriculum is related to the highest predicted performance. The combination of greater TOT and extended learning outside

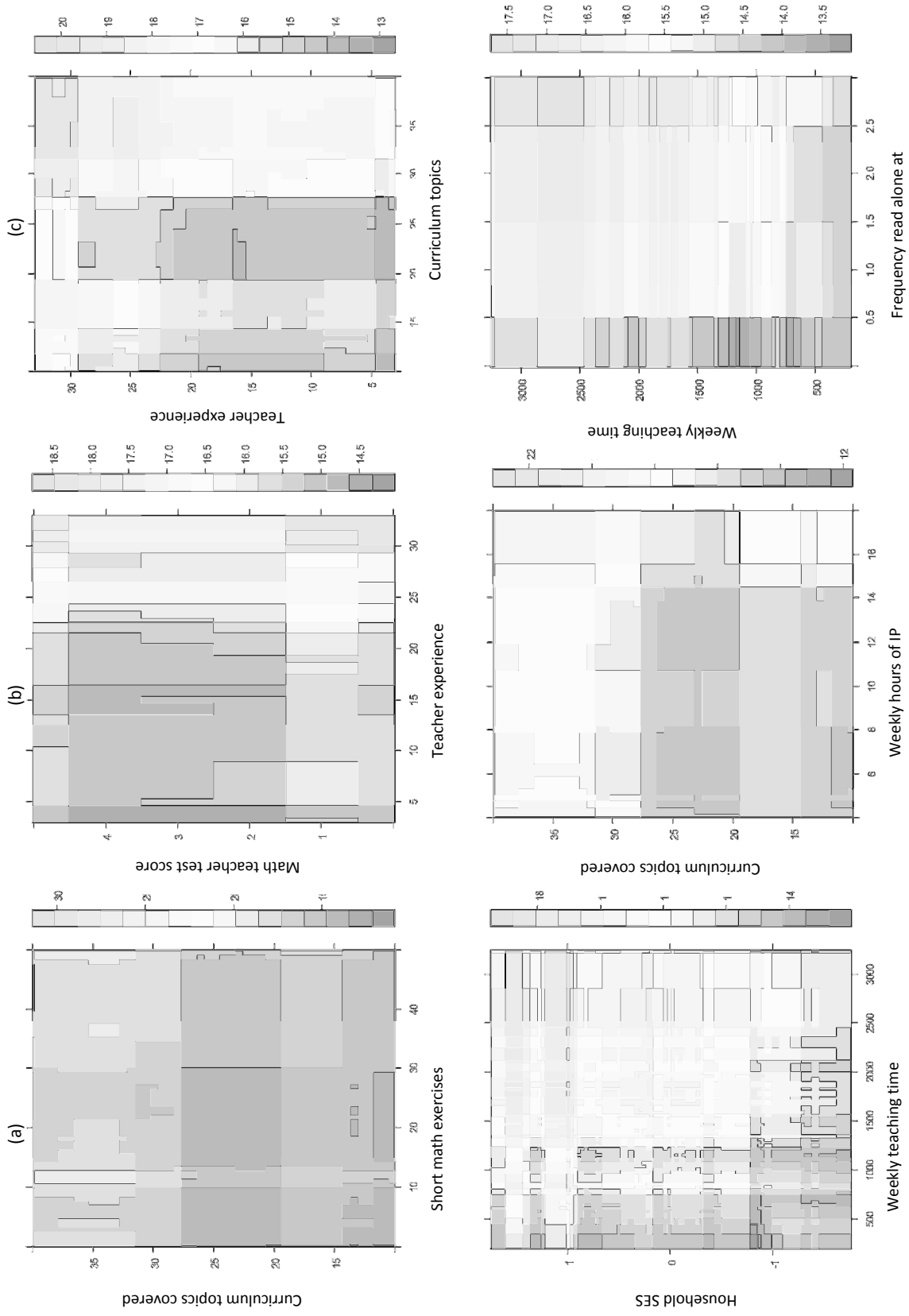
of the classroom in the form of independent reading also bring noticeable score gains (see figure 2.5f). Irrespective of teaching time, students who do not engage in any independent reading at home are predicted to perform worse than students who engage in some reading.

The pattern emerging from figure 2.5d indicates that irrespective of home affluence, students benefit from being taught by a teacher who reports formal teaching of at least 25 hours per week. However, predicted performance is still observed to increase with home wealth where formal teaching is above 25 hours a week. This result may be indicative of a more general pattern of the relationship between student home background and access to better school functioning that is afforded by a certain level of wealth. This may be truer at the extremes of home SES. Focusing on the group of students with SES 1 standard deviation either side of the mean, there is clear evidence of augmented math performance associated with formal teaching time that is in line with national education policy. In an assessment of educator workload, Chisholm et al (2005) found that only half of a teacher's work week was actually spent on teaching, with time-on-task becoming progressively shorter as the week progressed. Overcoming the constraints to achieving adequate TOT - such as teacher absenteeism - as well as ensuring that time on task is spent on productive opportunities to learn need to be addressed by former disadvantaged schools if performance is to be improved.

For brevity's sake, this paper will not discuss in detail the joint partial dependence plots for the literacy model as the findings largely agree with those of the numeracy model. Rather, a brief summary of the core findings is provided. A positive interaction between teaching time and teacher experience indicate that students taught by more experienced language teachers who maximise their time-on-task are predicted to score better than students taught by less experienced teachers who have formal contact time of less than 20 hours a week. All else equal, increasing time-on-task brings positive returns to performance. Opportunity to learn through the use of class exercises are also related to improved performance particularly when a combination of exercises (words, sentences and paragraphs) are used. For example, engaging students in exercises of isolated words or sentences that extend over more than ½ a page is not found be to conducive to learning if not combined with more complex writing exercises.[36] An interesting positive interaction occurs between the length of the school day and home language. Students who report speaking English sometimes at home who attend schools with an extended school day (over 7 hours long) are predicted to perform better than students with a similar exposure to English at home but attend a school with a typical school day of less than 5 hours. School day length also interacts with teacher experience such that the negative effect of lack of experience may be countered by more time spent in school.

[36] The number of paragraph (sentence) exercises less than ½ a page long negatively interacts with the number of paragraph (sentence) exercises longer than ½ a page, indicating a trade-off.

Figure 2.5: Joint partial dependence plots illustrating two-way interactions from Grade 4 numeracy model



Notes: own calculations using NSES 2008. Plots generated using the gbm package in R.

2.6 Robustness checks

2.6.1 Sensitivity to dropping observations

I test the robustness of the BRT model results through dropping random subsets of the full sample. As the model is developed at the individual (student) level, random subsets of schools are dropped rather than individual students. It might be expected that estimating a BRT model on a smaller sub-sample of the sampled schools will not yield similar results. This is due to the fact that spatial differences in performance across provinces may be related to differences in school functioning and resource provision. For example, a recent study of the impact of provincial boundary changes on school performance by Gustafsson and Taylor (2013) found that a school switch from the traditionally poor performing province of the North West to the Gauteng province was associated with an improvement in mathematics performance. In order to ensure congruency in the underlying sample when dropping schools, the spatial distribution of schools at least at the provincial level needs to be retained. This becomes difficult when close to two-thirds of all South African schools are located within three provinces and the sampling design of the NSES study was based on this distribution. Dropping large numbers of schools from the sample needs to take this into account.

Table 2.4 presents the ten most influential predictors in Grade 4 numeracy models developed for samples that exclude ten percent and 20 percent of schools compared to the model developed on the full sample. Nine of the ten most influential predictors from the full model are found to be similarly influential in the smaller sample models. Although there are some differences in ranking, school SES features as the most influential factor across all three models. The partial dependence plots of the predictors indicated in table 2.4 are very similar to those indicated in figure 2.3, expressing similar trends in inflection points and non-linearities.[37] The final column of table 2.4 therefore presents results from a BRT model built on a sub-sample that excludes 30 percent of schools selected randomly within provinces. The ten most influential predictors are identical to those from the full sample model with slight differences in ranking.

[37] Not shown here but available from the author on request.

Table 2.4: Most influential predictors across BRT models of numeracy score using sub-samples of the NSES (2008) data

	Whole sample	Dropping random 10% of schools	Dropping random 20% of schools	Dropping 30% of schools[a]
School SES	**8.08**	**8.59**	**10.6**	**10.7**
Curriculum topics covered	**7.87**	**7.39**	**7.09**	**6.14**
Short math exercises	**7.23**	**5.97**	**6.81**	**3.74**
Class size	**6.29**	**8.36**	**6.47**	**7.3**
School pupil-teacher ratio	**5.75**	**5.96**	**5.43**	**5.85**
Teacher's weekly teaching time	**5.15**	**4.5**	**7.36**	**4.69**
Household SES	**4.68**	**4.5**	**4.01**	**4.98**
Intermediate Phase math classes	**3.68**	**4.4**	**3.56**	**3.98**
Teacher experience	**3.53**	**4.3**		**3.55**
Age	**3.37**		**3.78**	**3.28**
Number of iterations	1950	1750	1100	1550
Shrinkage	0.1	0.1	0.1	0.1
Tree complexity	2	2	2	2
RMSE	6.49	6.55	6.51	6.43
R-squared	0.46	0.47	0.45	0.47
Observations	11894	10910	9389	8294
		minimum	maximum	

Notes: own calculations using NSES 2008

[a]Randomised within province.

2.6.2 Comparisons across independent data sets

As a further test of robustness, model results are compared across multiple data sets. In order to ensure comparability, a BRT model fitted to the Grade 5 NSES numeracy (2009) is compared to a BRT model fitted to the Grade 6 SACMEQ III numeracy scores (2007).[38] As former school department is not available in the SACMEQ III datasets, this split needs to be made using a proxy indicator. In the case of SACMEQ III the split is created using school SES given that the former black African department is highly correlated with school wealth, particularly amongst the poorest quintiles. Information for the Gauteng province is also excluded from the SACMEQ III dataset. However, comparability is not fully guaranteed given the different years of assessment and sampling designs as well as different test instruments used for measuring performance across the two datasets.[39]

[38] Similar analysis for the Grade 5 NSES (2009) literacy scores and Grade 6 SACMEQ (2007) literacy scores was conducted, but has been excluded from this paper. Results are available for the author on request.

[39] The numeracy test given to Grade 5 students in the 2009 wave of the NSES survey was comprised of questions set at grade levels 2 through 5 in line with the South African National curriculum.

The BRT models initially control for all available predictors that may or may not be similar across the datasets, referred to as model (a) in table 2.5. It is encouraging to find that at least a third of the predictors included in model (a) across datasets are identical. With regards to the remaining predictors, they appear to be indicative of the same underlying factors. For example, indicators of TOT and OTL emerge as important for determining numeracy test scores, as do indicators of the learning environment at home. Model (b) controls for only the 37 predictors that are common to the two datasets. Of the 10 most important variables reported, the top 7 are identical across datasets with slight differences in relative ranking. Partial dependence plots (not provided here) further indicate very similar patterns of the predicted relationships between these variables and test performance. This is despite differences in the scaling and distribution of the dependent variable across the two datasets as well as the fact that the two test instruments may be capturing different levels of numeracy proficiency.[40] This further illustrates the robustness of the BRT modelling approach.

2.6.3 Comparisons with alternative modelling approaches

Table 2.6 compares the predictive performance of the BRT model to the traditional linear least squares (LS) regression model adopted in education production modelling and other competing non-linear, non-parametric techniques. Predictive performance is assessed using the predicted R-squared and predicted root mean squared error generated from a 30 percent test sample that is held back as the model fitting stage. This allows us to determine whether or not the fitted model is capable of providing valid predictions for new observations. We would expect a lower R-squared and a more conservative (higher) RMSE from the test dataset, although dramatic differences in the training and test R-squared values may be symptomatic of model overfitting. All models represented in table 2.6 are built on the same training dataset. In order to avoid contaminating the holdout dataset,[41] model parameters were chosen using suggested defaults.

Although the predictive performance of the LS numeracy and literacy models (column 2) is shown to improve on that of a single regression tree (column 6),[42] it is substantially lower than that of the BRT models adopted for the earlier analysis of this paper. A generalised additive model (GAM) introduces flexibility into the general linear form of the LS model through estimating non-parametric functions - for example a cubic smoothing spline - that relates the covariates to the outcome of

[40] The distribution of numeracy test scores for the NSES data is skewed to the left with a larger variance whilst the SACMEQ test score data is normally distributed.
[41] A common usage for the hold-out (test) dataset is for training a model so as to determine the most suitable final model parameters. If the test data is repeatedly used for model selection purposes, then it may no longer provide an unbiased indication of the predictive error in the model.
[42] This reiterates the earlier statement that the capacity of a single regression tree for prediction is limited.

interest.[43] Whilst the predictive performance of the GAMs is a clear improvement over the LS models, they fall short of the BRT models.

Table 2.5: Most influential variables in BRT models of numeracy across the NSES Grade 5 (2009) and SACMEQ Grade 6 (2007) datasets

NSES Grade 5 numeracy (2009)			**SACMEQ Grade 6 numeracy (2007)**		
	(a)	(b)		(a)	(b)
School SES	**9.9**	**17.5**	School pupil-teacher ratio	**11.5**	**17.8**
Age	**7.2**	**4.2**	School SES	**10.2**	**18.1**
School pupil-teacher ratio	**6.4**	**12.2**	Help with reading homework	**8.3**	
Curriculum topics	**6.3**		Math teacher test score	**6.3**	
Class size (average)	**6.2**	**12.9**	Age	**5.1**	**7.9**
Frequency watch television in English	**4.9**		Household SES	**4.9**	**10.5**
Complex math exercises	**4.8**		Class size (average)	**4.3**	**8.9**
Short math exercises	**4.6**		Mother's education	**4.1**	
Household SES	**4.2**	**7.4**	Classroom resources	**3.9**	
Teacher experience	**4.2**	**8.7**	Teacher's teaching time	**3.8**	**9.6**
Long math exercises	**3.7**		Household chores	**3.7**	
Frequency read to by an adult	**3.4**		Reading teacher test score	**3.4**	
Frequency read alone at home	**3.2**		School head's experience	**3.3**	
Frequency of homework	**2.6**	**2.8**	Fax facilities at school	**2.9**	**3.3**
Weekly hours of IP math	**2.2**		Father's education	**2.5**	
Number of iterations	2900	1950	Number of iterations	2550	1800
Shrinkage	0.05	0.10	Shrinkage	0.025	0.05
Tree complexity	2	2	Tree complexity	2	2
Root mean squared error	7.44	6.77	Root mean squared error	3.06	3.12
R-squared	0.43	0.42	R-squared	0.40	0.38
CV root mean squared error	7.72	7.14	CV root mean squared error	3.38	3.33
CV R-squared	0.37	0.35	CV R-squared	0.32	0.29
				minimum	maximum

Notes: own calculations using NSES (2009) and SACMEQ III (2007). Model (a) incorporates all relevant predictors available from the survey instruments, whilst model (b) only includes those predictors that are common across the two surveys.

[43] A more detailed description of how GAMs are fit to data can be found in Hastie and Tibshirani (1990)

Table 2.6: Model performance of competing approaches using training and test data splits

	(1) BRT	(2) LS	(3) GAM	(4) RF	(5) BART	(6) Single RT	(7) RE-EM
				Grade 4 numeracy model			
Training RMSE	6.24	7.46	7.02	6.77	6.40	7.78	6.98
Training R-squared	0.46	0.23	0.32	0.37	0.43	0.18	0.38
Predicted RMSE	6.84	7.50	7.13	6.90	6.77	7.78	7.25
Predicted R-squared	0.37	0.23	0.30	0.35	0.37	0.14	0.34
				Grade 4 literacy model			
Training RMSE	5.58	6.35	6.12	6.02	5.69	6.93	5.98
Training R-squared	0.43	0.27	0.32	0.34	0.40	0.12	0.34
Predicted RMSE	5.93	6.38	6.19	6.07	5.95	7.06	6.14
Predicted R-squared	0.36	0.24	0.29	0.34	0.37	0.09	0.33

Notes: own calculations using NSES 2008. BRT = boosted regression tree, LS = least squares, GAM = generalised additive model, RF = random forest, BART = Bayesian additive regression tree, RE-EM = random effects expectation maximisation. Training and test sets are based on a 70:30 data split respectively. The R-squared values for the OLS and GAM models represent measures of the adjusted R-squared, whilst the R-squared values for the BRT, RF, BART and REEM models are calculated as the proportion of variance explained.

Like BRT modelling, random forest (RF) and Bayesian additive regression tree (BART) models fall within the group of ensemble methods; that is, the model is constructed as a collection of many individual regression trees. The iterative process of building a RF model combines bagging and randomized node optimisation (Breiman, 2001). A single tree is grown on a bootstrapped sample of size *N* where each node split is determined by recursively selecting the best variable from a subset of *m* variables chosen at random[44] until the minimum terminal node size is reached. This algorithm (known as "feature bagging") ensures that the final "forest" is made up of many different trees[45] that are less likely to be correlated in the likelihood that a covariate/s are very strong predictors of the outcome. In simulated and real data applications RFs have been shown to achieve RMSE values as low as boosting (Hastie et al., 2009). BART modelling is similar in spirit to BRT except that it uses a prior[46] instead of bagging and shrinkage to weaken the contribution of each individual tree to the final prediction and a Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm is used to fit the sum-of-trees model (Chipman, George & McCulloch, 2012). The "randomForest" and "BayesTree" libraries in R are used to fit the RF and BART test score models respectively.

The results of these regression tree techniques (columns 4 and 5 of table 2.6) yield RMSE and R-squared values for the hold-out data that are similar in magnitude to that of the BRT models.

[44] The default is $\sqrt{p}$ where p is the total number of covariates eligible for selection.
[45] The default number of trees is 500.
[46] The priors are set for the regression tree parameters, specifically the tree size (depth) and the parameter values associated with the terminal nodes, as well as the error variance (see Chipman et al. (2012) for further discussion). The number of trees to be used in the final model also needs to be chosen, with the default choice being 200.

However, differences in the magnitude of these performance measures across the training and test datasets indicate that the BRT and BART models may be overfitting the test score data. Correlation of the fitted response values for the BRT, BART and RF models to the observed response values returns the strongest positive correlation for the RF models (in excess of 0.9) followed by the BRT model (approximately 0.7). It is therefore worthwhile investigating whether or not the "simpler" RF model provides different results to the BRT model.

Variable importance for random forests can be constructed in exactly the same way as for boosting. However, the most widely adopted measure of importance in random forest models is the increase in MSE that occurs from random permutation of a given variable in the out-of-bag (OOB) samples. This permutation score has been shown to be a more reliable measure of importance as it is less likely to be biased in favour of predictors with many values (see Strobl & Boulesteix, 2007). The variable importance for numeracy and literacy scores using BRT and RF models are shown in table 2.7. It is interesting to note that there is far more similarity between the BRT importance ranking and the RF permutation-importance measure (RF_1) than the RF importance measure that is calculated identically to the BRT model (RF_2). In the case of the literacy model, 12 of the 15 most influential variables are found to be equivalent across the two modelling approaches, and similarly 14 of the 15 most influential variables for the numeracy model. Partial dependence plots generated for the variables included in table 2.7 (results not shown here) indicate almost identical associations with the fitted response as was observed for the BRT models. Random forests therefore appear to be a viable alternative to boosting for modelling education production.

2.6.4 Regression Tree Modelling with Clustered Data

The final column of table 2.6 presents the results from a mixed-effects tree-based approach by Sela and Simonoff (2011). "Mixed-effects" refers to the use of both fixed and random effects in the same analysis. Mixed effects modelling has a natural application to nested data, such as students within schools, as it makes provision for the explicit modelling of a variety of correlation patterns in the data; for example, within-school correlation in errors. Specifically, a random intercepts model is estimated which includes a fixed function that relates schooling inputs to the test score. The estimation process begins by estimating a regression tree assuming zero random effects. The random effects are then estimated assuming that the fitted regression tree from the first stage is correct. This process is repeated until the random effects converge, similar to the Expectation-Maximisation (EM) algorithm of Dempster, Laird and Rubin (1977). It is for this reason that Sela and Simonoff (2011) refer to this model as a Random-Effects/EM (RE-EM) Tree. Unlike ensemble methods, the final RE-EM model predictions are based on only one regression tree (the fixed part of

the model). The "REEMtree" library in R is used to fit the models (Sela & Simonoff, 2011).Despite the relative simplicity of the RE-EM model, the predictive performance is comparable to that of the ensemble method approaches as evidenced by similar predicted R-squared and RMSE values. Figure 2.6 depicts the final fitted RE-EM regression tree (only the Grade 4 numeracy test score model is shown) where predictions for each terminal node are indicated by bold boxes. Twelve of the fifteen variables used for splitting in the RE-EM tree are identical to influential variables identified in the BRT and RF numeracy models.

Table 2.7: Variable importance across boosting and random forest models of numeracy and literacy

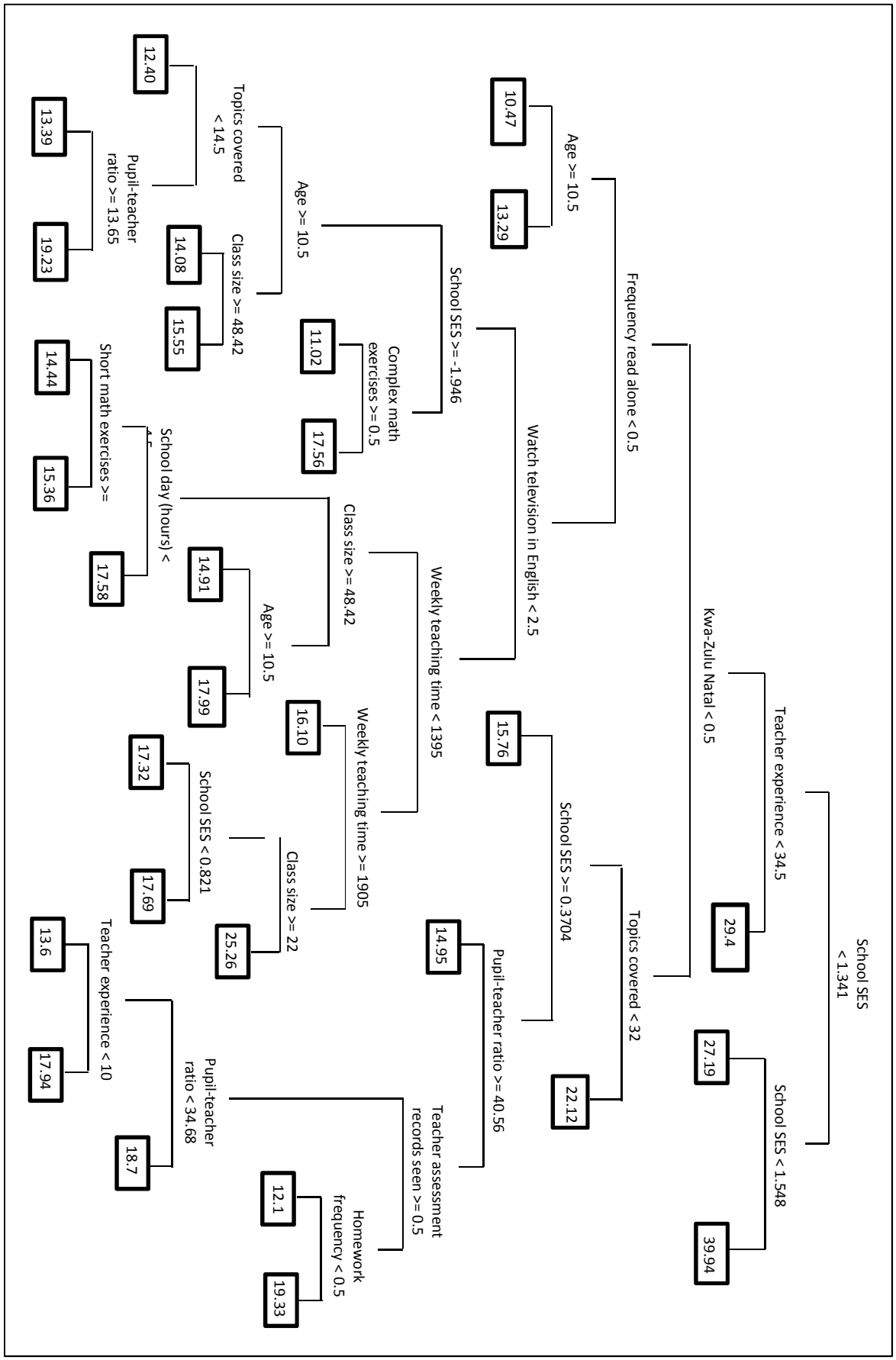
Grade 4 literacy				Grade 4 numeracy			
	BRT	RF_1	RF_2		BRT	RF_1	RF_2
School SES	**11.1**	**2.18**	**6.07**	School SES	**8.08**	**2.74**	**5.31**
School pupil-teacher ratio	**6.76**	**2.57**	**2.81**	Curriculum topics covered	**7.87**	**2.29**	**3.4**
Household SES	**5.91**	**2.29**	**9.48**	Short math exercises	**7.23**	**1.98**	**2.28**
Class size	**5.75**	**2.31**	**2.37**	Class size	**6.29**	**2.34**	**2.96**
Teacher experience	**5.73**	**1.98**	**2.24**	School pupil-teacher ratio	**5.75**	**2.20**	**2.41**
Sentence writing more than ½ page	**5.6**		**2.41**	Teacher's weekly teaching time	**5.15**	**2.12**	**2.41**
Age	**4.7**	**2.67**	**5.18**	Household SES	**4.68**	**2.74**	**9.47**
Frequency watch television in English	**3.8**	**2.40**	**3.97**	Intermediate Phase math (weekly hours)	**3.68**		
Teacher's weekly teaching time	**3.66**	**1.92**		Teacher experience	**3.53**	**2.28**	**3.83**
Word exercises less than ½ page	**3.46**	**2.32**	**1.92**	Age	**3.37**	**2.69**	**5.28**
Paragraph exercises less than ½ page	**2.65**		**1.95**	Complex math exercises	**3.12**	**1.58**	
Female	**2.42**	**1.70**	**2.33**	Long math exercises	**3.03**	**1.79**	
Word exercises more than ½ page	**2.34**			Frequency of reading homework	**2.69**	**2.38**	**3.96**
Frequency read alone at home	**2.27**	**1.96**	**3.92**	Kwa-Zulu Natal	**2.63**	**3.17**	**2.06**
Frequency of reading homework	**2.26**	**1.95**	**3.56**	Frequency read alone at home	**2.31**	**2.27**	**4.16**
Math teacher test score		**1.64**		Frequency watch television in English		**2.36**	**3.83**

minimum maximum

Notes: own calculations using NSES 2008. RF_1 calculates variable importance using OOB permutation. RF_2 calculates variable importance using the same indicator as in BRT.

The tree is quite complex (depth of 9), yet overall the results agree with the main findings of sections 2.5.3 and 2.5.4. Not accounting for random school effects, students attending high SES schools (1.5 standard deviations above average) have the highest estimated performance. Students taught by very experienced teachers are also predicted to have augmented performance, as are students taught in small class environments (less than 22 students) by a teacher whose weekly teaching time is within the expected bandwidth of 25 to 35 hours; this is provided that the student is engaged in reading and exposed to English regularly at home. It is evident from the graph that home background and school environment, particularly at the classroom level, interact to determine math-

Figure 2.6: RE-EM regression tree for Grade 4 numeracy



-ematical proficiency. For example, frequent contact with the test language and being of the correct age for the grade (younger than 11 years old) generally appears to be related to better numeracy performance regardless of the school and class contexts. Yet in spite of these factors, there are gains to being taught in small classes and being exposed to low pupil-to-teacher ratios and extended hours of learning.

2.7 Concluding Remarks

The primary aim of this paper was to propose machine learning techniques generally and ensemble methods specifically as alternative modelling approaches to the linear regression education production function. Whilst these methodologies have largely been reserved for prediction (as opposed to explanatory and/or causal investigation) I would agree with Shmueli (2010) that "neglecting to include predictive modelling and testing alongside explanatory modelling is losing the ability to test the relevance of existing theories and to discover new causal mechanisms". Machine learning techniques allow for more effective modelling of complex relationships and, as stated by Varian (2014), may be "the most natural for economic applications". Even if the results of predictive models limit the ability of researchers to make causal claims, they may assist in estimating the causal effect of an intervention should it occur. For example, Hill (2011) and Hill and Su (2013) show that the BART method may be used for causal inference in observational studies. Furthermore, with increased access to large datasets, advances in data collection and storage as well as the development of sophisticated modelling software comes a unique opportunity for researchers to bridge the gap between the development of new methodologies and their application in practice.

In light of the above findings, it would also be ideal to incorporate random effects into an ensemble method such as boosting or random forests. Hajjem, Bellavance and Larocque (2014) provide an extension of random forests to clustered data through an iterative method called "mixed-effects random forests" (MERF) that is similar in spirit to RE-EM except that the RT is replaced with a forest of RTs. A simulation study of 12 different data generating processes finds the lowest predicted mean square error for the MERF method when compared to a mixed-effects regression tree and a traditional regression tree. When compared to a RF the out-of-sample performance of MERF relative to RF is variable, although MERF is found to be more appropriate in contexts where the random effects are non-negligible

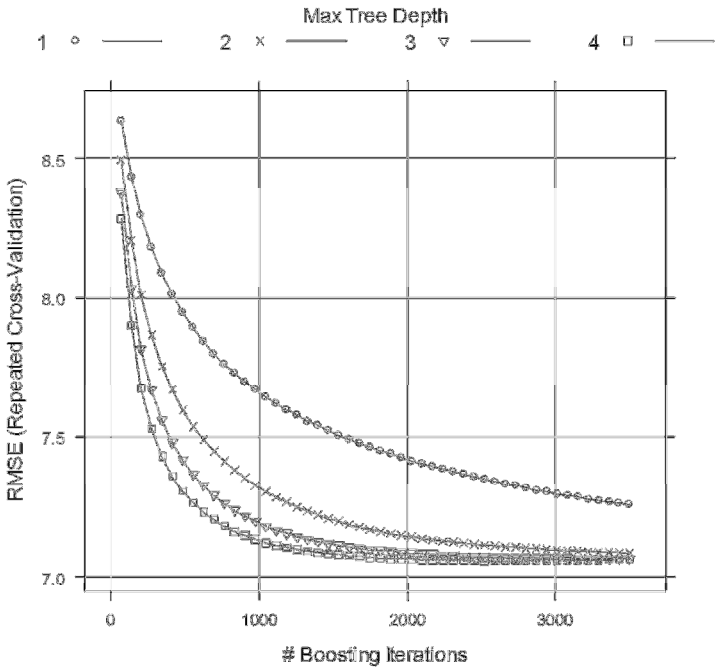
In this paper an attempt was made to model the education production process of disadvantaged schools in South Africa in an attempt to better understand the general ineffectiveness with which these schools transform schooling inputs into performance. Attempts were also made to identify those factors associated with augmented test scores that may be targeted by education

policy. The NSES dataset that contains data on student, household and school level characteristics as well as identifies former school department was employed. A boosted regression tree model that explicitly allowed for multi-level interactions between predictors was built for the Grade 4 numeracy and literacy test scores for the former DET and Homeland school sample. The results indicate that social contexts continue to be relevant for determining student outcomes. Tikly (2011:11) argues that a deeper appreciation of context is required in order to characterise good quality education, as it "encourages policy makers to take cognisance of changing national development needs, the kinds of schools that different students attend and the forms of educational disadvantage faced by different groups of students when considering policy options". The right blend of enabling processes/inputs at the level of national policy, the school and the home/community is vital for achieving the desired schooling outcomes. Less affluent South African schools face constraints – both real and perceived – that inhibit effectiveness, as "where communities are poor, have few material resources, and do not speak the language of instruction in their homes, there are few options to supplement the quality of teaching and learning in their schools" (Christie et al, 2007: 101).

The most significant positive interventions for the black school system would therefore be those which affect enabling inputs and processes, and work to overcome the gaps that often exist between schools, households/communities, and national policy. This includes the professional development of teaching staff in general, and principals specifically, to understand, choose, develop and evaluate relevant, effective practices within the context of their own school's status and culture. The encouragement and strengthening of parent involvement by principals is also vital. In bridging the learning gap that exists between the school and home environments, a better understanding of those classroom processes that disproportionately advantage poor students is required. This may include extending the amount of in-school learning time for children who lack the necessary supporting inputs at home.

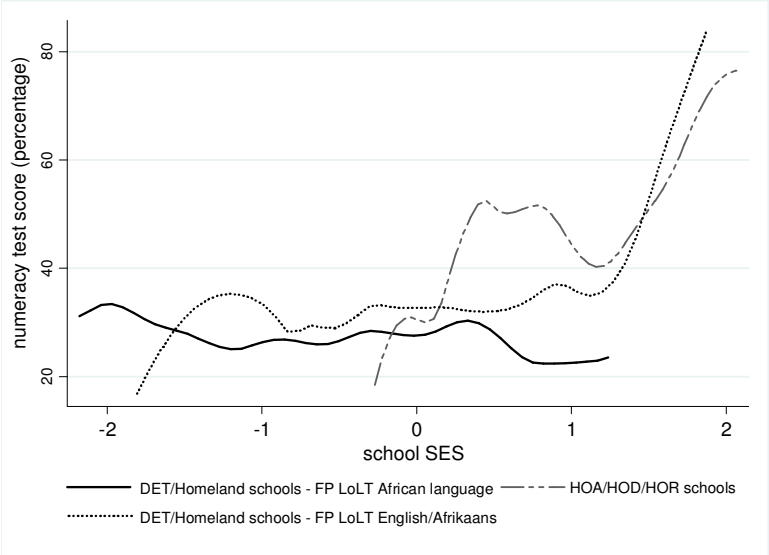
Appendix to Chapter 2

Figure A2.1: Relationship between RMSE and number of trees for numeracy score models fitted with three levels of tree complexity



Note: own calculations using NSES 2008. Model tuning conducted using caret and gbm packages in R. Due to computational constraints, parameter tuning does not incorporate bagging. A learning rate of 0.1 was used in all models.

Figure A2.2: Socio-economic gradients across former school department groupings



Note: own calculations using NSES 2008. SES gradients are fitted using kernel-weighted local polynomial smoothing. FP stands for foundation phase (Grade R – Grade 3) and LoLT stands for the language of learning and teaching. DET = Department of Education and Training, H = homeland, HOA = House of Assembly (white), HOD = House of Delegates (Indian), HOR = House of Representatives (coloured).

Table A2.1: Description of model covariates for NSES grade 4 numeracy and literacy models

Variable	Description	Type
Student age	Age of student in years	Continuous
Home language	Indicators of home language of student	Dummy
Frequency reads at home alone	Never = 0; once a week = 1; 2-3 times a week = 2; >3 times week = 3	Categorical
Living arrangement	Indicators of living with both parents, mother only or orphan	Dummy
Adult reads at home	Indicator of whether an adult reads at home	Dummy
Frequency read to by an adult	Never = 0; once a week = 1; 2-3 times a week = 2; >3 times week = 3	Categorical
Frequency of homework	Never = 0; once a week = 1; 2-3 times a week = 2; >3 times week = 3	Categorical
Help with homework	Indicator of receiving help from mother, father or sibling	Dummy
Frequency of speaking English at home	Never = 0; once a week = 1; 2-3 times a week = 2; >3 times week = 3	Categorical
Frequency of watching television in English	Never = 0; once a week = 1; 2-3 times a week = 2; >3 times week = 3	Categorical
Frequency of listening to the radio in English	Never = 0; once a week = 1; 2-3 times a week = 2; >3 times week = 3	Categorical
Household socio-economics status (SES)	Asset index (standardised): mean = 0 and s.d. = 1	Continuous
Number of children in the home	Indicators of only child, 1-2 siblings and >2 siblings	Dummy
School facility availability and functionality	Indicators of electricity, water, storage room, library, science laboratory, administration office and toilets. 0 = not present; 1 = present but non-functional; 2 = present and functional	Categorical
School technology availability and functionality	Indicators of telephone, copier, fax, internet and computers. 0 = not present; 1 = present but non-functional; 2 = present and functional	Categorical
School pupil-teacher ratio	School size / total teachers	Continuous
Library books	Indicator that students may take library books home	Dummy
Teacher absenteeism	Indicator that no teachers were absent on the day of the survey	Dummy
School poverty quintile	School wealth quintiles 1 through 5	Categorical
Timetable	Indicator that school timetable was seen.	Dummy
Length of typical school day (hours)	1 = <5 hours; 2 = 5-5.5 hours; 3 = 5.5-6 hours; 4 = 6-7 hours; 5 = >7 hours	Categorical
School socio-economic status	Average Socio-economic status of student sample in school (standardised): mean = 0 and s.d. = 1	Continuous
Functional sports facilities	0 = not present; 1 = present and non-functional; 2 = present and functional	Categorical
Teacher register	Indicator of teacher register available and up-to-date	Dummy
Learner teacher support materials	Indicators of LTSMs seen, up-to-date and used	Dummy
English and mathematics textbooks	Indicator of textbook availability for all students	Dummy
Textbooks in LoLT	Indicator of textbooks in language of instruction	Dummy

Table A2.1 continued: Description of model covariates for NSES grade 4 numeracy and literacy models

Variable	Description	Type
Teacher subject test score	Mathematics teacher test 0-5 marks; Language teacher test 0-7 marks.	Continuous
Average grade 4 class size	Total students in Grade 4 ÷ number of Grade 4 classes	Continuous
Teacher experience	Total years in teaching	Continuous
Teacher curriculum	Indicator that teacher has own copy of national curriculum	Dummy
Teacher's weekly instruction time (minutes)	Time devoted to in-school teaching across all phases (foundation, intermediate and senior)	Continuous
Teacher assessment records	Indicator that assessment records were seen	Dummy
Time spent on Intermediate Phase teaching per week (hours)	Total time devoted to subject instruction across Grades 4-6	Continuous
Mathematics curriculum topics covered	Count of topics covered through a comparison of student workbooks to national curriculum outline	Continuous
Number of mathematics exercises completed	As appearing in student workbooks. Divided into short exercises (less than 5 lines), long exercises (more than 5 lines) and complex exercises.	Continuous
Number of language exercises completed	As appearing in student workbooks. Divided into paragraph, word and sentence exercises as well as shorter than ½ a page and longer than ½ a page.	Continuous

Notes: Household socio-economic status (SES) was generated using first principal component analysis of availability of 10 items in the household including electricity, tap water, toilet, car, computer, refrigerator, washing machine and daily newspaper. Primary schooling is sub-divided into three phases: intermediate phase covers Grades R to 3; foundation phase covers Grades 4 to 6; and senior phase covers Grades 7 (which extends to Grades 8 and 9 in secondary schooling). LoLT stands for the language of learning and teaching.

Table A2.2: Variable correlations

Numeracy model		
Variable 1	Variable 2	ρ_{12}
Frequency adult reads to student	Adult reads at home	0.0000*
Short math exercises	Curriculum topics covered	0.1944
Teacher experience	Math teacher test score	0.1835
3 or more children in the home	Western Cape province	0.0000*
Curriculum topics covered	Teacher experience	0.0478
Frequency adult reads to student	Western Cape province	0.0000*
Weekly teaching time	Household SES	0.1555
LOLT textbooks for all students	Curriculum topics covered	0.1034
Curriculum topics covered	Hours of IP math per week	0.2175
Weekly teaching time	Frequency student reads at home	0.0630
Literacy model		
Variable 1	Variable 2	ρ_{12}
Frequency adult reads to student	Adult reads at home	0.0000*
Staff computers present and functional	North West province	0.0000*
School poverty quintile	Help with homework from father	0.0000*
Zero teachers absent	Electricity present and functional	0.0020
Weekly teaching time	Teacher experience	0.0022
Home language African	3 or more children in the home	0.0000*
Toilets present and functional	Only child in the home	0.0000*
Shortage of LTSM	Permanent principal	0.0000*
LTSM unused	Speak English regularly at home	0.5730
3 or more children in the home	North West province	0.0050*

Notes: own calculations using NSES 2008. *Pearson's chi² test is used in the case of correlations between two binary variables and the p-value reported (shown in italics).

Chapter 3

A question of efficiency: decomposing South African reading test scores using PIRLS 2006

This paper assesses the PIRLS (2006) reading score gap observed between the historically advantaged and historically disadvantaged school systems in South Africa. The methodology employed by this paper builds on the work of Botezat and Seiberlich (2013) that addressed the issues with the Oaxaca-Blinder decomposition for analysing achievement gaps. The methodological contribution of this paper uses the reweighting decomposition technique of DiNardo (2002) to identify two separate "treatments" of attending a better school system. Estimates indicate that policy directed at improving school efficiency and school resourcing within the historically disadvantaged school system could as much as halve the average performance gap.

3.1 Introduction

Under the apartheid government, resources for black African schools were centrally controlled by a Department of Education and Training (DET), with the control of white, Indian and coloured schools assigned to separate bodies.[47] This system led to the creation of a highly inequitable distribution of school resources across both racial and regional lines, resulting in large discrepancies in the educational attainment and performance of the different education systems. Despite concerted efforts to equalise the distribution of school resources in the South African education system, a large portion of the system, primarily historically black African schools, still fails to provide quality basic education (Van der Berg et al., 2002; Van der Berg, 2007; Van der Berg, 2008). This is confirmed by the weak performance of South African students on international tests, even when compared to countries with comparatively resource-poor education systems. The "bimodal" pattern of test results that is typically observed illustrates how far historically Black schools continue to lag behind white, Indian and coloured schools in performance and that different data generating processes exist for historically white schools than for historically black African schools (see for example Gustafsson, 2007; Fleisch, 2008; Taylor, 2011; Shepherd, 2013; Spaull, 2013).

[47] House of Assembly for white schools, House of Representatives for Indian schools and House of Delegates for coloured schools.

Recent studies have made divergent conclusions. In a cluster fixed effects analysis of schooling attainment using the first wave of the National Income Dynamics Study (NIDS) dataset, Timæus, Simelane and Letsoalo (2013) argue that the poor attainment and low matriculation success of disadvantaged, mostly black African, students is not due to the poor performance of the former black African school system, but rather can be accounted for by home/parent background and socio-economic status. Although the link between race and performance is strong, black African children from better socio-economic backgrounds perform exceedingly better than their less-affluent counterparts. Socio-economic status and parent education are found to be significant in explaining the variation in performance results (Taylor & Yu, 2009; Van der Berg, 2008) and attainment (Timaeus & Boler, 2007; Lam, Ardington & Leibbrandt, 2011; Timæus et al., 2013). However, these variables are most likely positively related to unobservable home background characteristics that are themselves related to school choice such as the value that parents place on education.

This paper aims to shed light on the source/s of discrepancy in performance between former black African and former advantaged schools, and whether the discrepancy comes as a result of differences in school quality[48] or access to a lower level of (quality) resources. I use data from the 2006 Progress in International Reading Literacy Study (PIRLS) to decompose the performance gap between those schools that tested in English or Afrikaans (as a proxy for the historically advantaged school system) and those schools that tested in an African language (as a proxy for the historically disadvantaged black African school system). Traditionally, regression decomposition techniques such as that of Oaxaca (1973) and Blinder (1973) are employed to disentangle the distributional effects of educational input endowment (explained effect) from that of the returns to these inputs (unexplained effect), with much of the emphasis falling on the relative size of the former.

A recent study by Botezat and Seiberlich (2013) employs a semiparametric approach to decomposing performance gaps in Eastern European countries. Their construction of a counterfactual mean using propensity score matching allows assessment of the extent to which differences in student and home background characteristics contribute to explaining the observable gaps in school performance (explained gap), with the remaining gap due to differences in schooling systems. Construction of a counterfactual in this way is important as recent papers have confirmed that the functional form assumptions of the parametric Oaxaca-Blinder decomposition can give misleading results (Barsky, Bound, Charles & Lupton, 2002; Mora, 2008). Furthermore, as will be discussed, the unexplained component constructed in this way is more representative of the average treatment effect of attending a school within a particular school system.

[48] School quality is defined as the extent to which a school and its constituent parts (teachers, management, culture and infrastructure) improve a student's learning.

Whilst the semiparametric approach of Botezat and Seiberlich (2013) employs propensity score matching, the analysis of this paper adopts the reweighting approach of DiNardo (2002) and DiNardo, Fortin, and Lemieux (1996) to construct the counterfactual of interest. This approach allows the unexplained performance gap to be separated into two "treatments" of attending a different school type. The first of these is the effect of attending a school within a school system that offers higher returns to educational inputs, or school efficiency gap. The second component of the unexplained gap is due to differences in the distribution of school resources across the two school systems, or school resource gap. The author proposes that these two components of the unexplained gap provide education policy with two different tools for assessing how the performance gap between two students attending schools within different school sub-systems might be closed.

The remainder of the paper is laid out as follows: Section 3.2 focuses on the identification strategy and decomposition methodology adopted by this study. Section 3.3 introduces the PIRLS 2006 data and descriptive statistics for the two school groups under comparison. Section 3.4 discusses the empirical results. Section 3.5 concludes and outlines policy recommendations informed by the findings.

3.2 Methodology

3.2.1 Oaxaca-Blinder decomposition

Decomposition methods, beginning with the seminal work of Oaxaca (1973) and Blinder (1973) and the Oaxaca-Blinder (OB) decomposition, finds its roots in the labour economics literature. Its adoption in the context of educational outcomes is fairly recent, with studies chiefly emanating from the education production function literature with attempts made to determine the extent to which performance gaps may be explained by differences in student and school characteristics, with the remaining (unexplained) gap due to differences in the quality or effectiveness of the different education processes. Applications exist across geographical lines (Tansel, 1999; Ammermueller, 2006; McEwan, 2008; Burger, 2011; Botezat & Seiberlich, 2013), school types (Krieg & Storer, 2006; Duncan & Sandy, 2007), across time (Barrera-Osorio, 2011; Cattaneo & Wolter, 2012; Da Maia, 2012; Sakellariou, 2012) and across race and gender (Sohn, 2012a, 2012b).

Two of the most important developments in the decomposition methodology literature are (i) extensions to the entire outcome distribution (Juhn, Murphy & Pierce, 1993; DiNardo et al., 1996) and (ii) linkages to the treatment effect literature. Recent contributions by Barsky et al (2002), Fortin, Lemieux and Firpo (2011) and Słoczyński (2014) have shown that the OB decomposition

provides a consistent estimator of the population average treatment effect of the treated (PATT). Kline (2011) has further shown the method to be equivalent to a propensity score reweighting estimator that is based on a linear model for the odds of being treated, making it a "doubly robust" estimator of the counterfactual mean.[49]

I consider a population of N students indexed by $i = 1, \dots, N$ that are divided into two mutually exclusive groups denoted by the binary variable g_i where $g_i = 0$ represents membership to the group of historically disadvantaged schools (control group) and $g_i = 1$ represents membership to the group of historically advantaged schools (treatment group). The outcome of interest is the reading test score Y_{ig} and we further observe a set of k controls X_i. As in the treatment effect literature, Y_{i0} and Y_{i1} can be interpreted as two potential outcomes for student *i*. While both of these outcomes are observed, only one is realized, with the realized outcome given by:

$$Y_{ig} = Y_i(0)(g_i - 1) + Y_i(1)g_i \qquad [3.1]$$

The OB model is based on a linear model for the potential outcomes that allows for divergent regression coefficients across the two groups:

$$Y_{ig} = \beta_g{}'X_{ig} + \varepsilon_{ig} \quad \text{where } E[\varepsilon_{ig}|X_i, g_i] = 0 \quad \text{for } g\epsilon\{0,1\} \qquad [3.2]$$

Given [3.2], there are three possible reasons why the distribution of reading scores between the two school types could differ: i) differences between the returns structures β_0 and β_1; ii) differences in the distribution of observable characteristics X; and iii) differences in the distribution of unobservable characteristics ε. The aim of decomposition is to separate the contribution of (i) from (ii) and (iii).

In order for the decomposition to follow a partial equilibrium approach, I restrict the counterfactual returns structure to one of a "simple" counterfactual treatment in that the only alternative state of the world for group A would be the returns structure faced by group B, and vice versa.[50] Knowledge of β_0 and β_1 allows us to compute a simple counterfactual of this type; for example, "what would be the distribution of reading scores for students in group 0?", and vice versa.[51] Given this counterfactual, I am able to decompose the mean difference in the performance

[49] If the true odds-of-treatment are linear, then the Oaxaca Blinder estimate of the average treatment effect will be identified even if the model for potential outcomes is misspecified, provided that unconfoundedness and overlap hold. Conversely, if the model for potential outcomes is correct, the Oaxaca Blinder estimate will identify the average treatment effect even if overlap fails and/or the implicit model for the odds of treatment is incorrect.

[50] This rules out the existence of some other counterfactual returns structure that would prevail if, for instance, students from advantaged schools were no longer enrolled in those schools.

[51] The choice of whether to construct the counterfactual from the returns structure of group 1 or 0 corresponds to two methods of decomposing the differences in student characteristics (Krieg & Storer, 2006: 569). The research question posed by this study favours the use of group 1 returns structure in order to calculate the counterfactual

of students in school type 0 and those in school type 1 into a component attributable to differences in the observed characteristics of students and their schools (explained component) and a component attributable to differences in the returns structure to these characteristics (unexplained component); that is:

$$\begin{aligned} E[Y_i|g_i = 1] - E[Y_i|g_i = 0] &= \hat{\beta}_1'^{\bar{X}_1} - \hat{\beta}_0'^{\bar{X}_0} \\ &= \hat{\beta}_1'\bar{X}_1 - \hat{\beta}_1'\bar{X}_0 + \hat{\beta}_1'\bar{X}_0 - \hat{\beta}_0'\bar{X}_0 \\ &= \bar{X}_0'(\hat{\beta}_1 - \hat{\beta}_0) + (\bar{X}_1 - \bar{X}_0)'\hat{\beta}_1 \\ &= \hat{\Delta}_S + \hat{\Delta}_X \end{aligned} \quad [3.3]$$

where $\hat{\Delta}_S$ represents the unexplained component of the wage gap and $\hat{\Delta}_X$ represents the explained component.

From Słoczyński (2014), the unexplained component of the OB decomposition in [3.3] can be shown to represent the average treatment effect of the untreated (PATN) as follows:

$$\begin{aligned} &E[Y_i|g_i = 1] - E[Y_i|g_i = 0] \\ &= E[X_i|g_i = 0]'(\beta_1 - \beta_0) + (E[X_i|g_i = 1] - E[X_i|g_i = 0])'\beta_1 \\ &= E[Y_i(1) - Y_i(0)|g_i = 0] + \{E[\mathrm{Y}_i(1)|g_i = 1] - E[\mathrm{Y}_i(1)|g_i = 0]\} \\ &= \tau_{PATN} + \{E[\mathrm{Y}_i(1)|g_i = 1] - E[\mathrm{Y}_i(1)|g_i = 0]\} \end{aligned} \quad [3.4]$$

The second component of [3.4] represents the extent to which the control group (0) and treated group (1) are on average different, that is, the "selection bias".[52] The assumption of simple counterfactuals severely limits the interpretation of the unexplained component as a causal effect. It must therefore be made clear that whilst this paper makes reference to concepts of "effect" and "treatment", no claims of causality are made.

As in the treatment literature, further assumptions need to be made in order to identify the PATN. The first of these is ignorability (unconfoundedness) which states that the distribution of unobservable determinants of test performance are the same across both groups after controlling for observable characteristics; that is, $g_i \perp Y_{i0}, Y_{i1}|X_i$, ruling out selection into group 1 or 0 based on unobservables. Secondly, I assume overlapping support in that there do not exist any (sets of) values of X which would perfectly predict membership to either group 0 or 1. It is plausible that parents may select the schools which their children attend. If this is the case, differences in student body

distribution as we ask the question: what if students attending historically black African schools received the same treatment as students attending historically advantaged schools, and if so, what would the gap in reading scores be?

[52] Similarly, choosing the returns structure of disadvantaged schools as the counterfactual, we can decompose the test score gap into the average treatment effect on the treated (PATT) and selection bias; that is, $\tau_{\mathrm{PATT}} + \{\mathrm{E}[\mathrm{Y_i}(0)|\mathrm{g_i} = 1] - \mathrm{E}[\mathrm{Y_i}(0)|\mathrm{g_i} = 0]\}$

composition would not be wholly exogenous and the conditional distribution of $X, \varepsilon | g = 1$ may be different from the distribution of $X, \varepsilon | g = 0$. The conditional independence assumption does not necessarily rule out the possibility that these distributions may be different, but it constrains their relationship. Specifically, the joint densities of X and ε for groups 0 and 1 have to be similar up to a ratio of conditional probabilities (Fortin et al, 2011).[53]

3.2.2 Semi-parametric decomposition

Recent papers have revealed that the assumption of a linear condition mean function of the traditional parametric OB decomposition can give quit misleading results. Barsky et al (2002) show that the unexplained and explained components will be inconsistently estimated if the conditional mean function is truly non-linear. A further criticism of the OB decomposition is that it ignores issues of common support. This is confirmed by the fact that, until recently, all attempts to decompose student performance gaps have made no reference to issues of either functional form or overlap. This is in stark contrast to a substantial proportion of labour market applications of OB decomposition over the past decade where much thought has been devoted to understanding the implications of incorrect functional form and lack of overlap (DiNardo et al., 1996;Barsky et al., 2002; J DiNardo, 2002; Lemieux, 2002; Mora, 2008; Ñopo, 2008). Furthermore, not all covariates contained in X can be considered as pre-treatment variables and, as a consequence of treatment, may assume different values across the two school groups (Schneeweis, 2011).[54] When X is affected by treatment, the unexplained component will represent a partial effect of treatment that is netted from the indirect effect of treatment through changes in X (Fortin et al, 2011).

In their analysis of performance gaps across eight European countries, Botezat and Seiberlich (2013) were the first within the educational production literature to apply a semi-parametric OB decomposition approach that accounts for these issues. Specifically, their estimates of the counterfactual means are identified by matching on propensity scores (as in Frölich, 2004). In this way, the counterfactuals are constructed for individual students who are actually comparable. Propensity score reweighting as laid out in DiNardo et al (1996) and DiNardo (2002) is another popular technique for constructing counterfactuals. Botezat and Seiberlich's (2013) choice of matching over other techniques for constructing counterfactuals is driven by Frölich's (2004) findings

[53] The literature offers several solutions to deal with violation of the conditional mean independence assumption, the traditional methods being the use of a control function (Heckman, 1979; Heckman & Robb, 1985) or instrumental variable models (Heckman and Vytlacil, 2001, 2005). Arguably the best way of dealing with selection and endogeneity is to use panel data methods. Given the cross-sectional nature of the data employed by this study, we need to be mindful of potential bias in the model parameters when interpreting the results.

[54] Furthermore, as observed by Botezat and Seiberlich (2013: 736), some school resources are perfect predictors of treatment assignment and therefore cannot be included as controls in the propensity score model.

that reweighting performs considerably worse than matching when estimating average treatment effects. However Busso, DiNardo and McCrary (2014) show that, unlike un-normalized reweighting as considered by Frölich (2004), *normalized* reweighting compares favourably with matching, except in cases of sufficiently low overlap where matching is more effective. The identification strategy undertaken in this paper makes use of normalized reweighting. I will proceed by first discussing the identification strategy adopted by this paper, after which some relevant comparisons with the approach of Botezat and Seiberlich (2013) will be made.

Specifically, let student *i* in the group of historically advantaged schools (g = 1) have weight $w_i = \frac{\Pr(X_i|g=0)}{\Pr(X_i|g=1)}$, where w_i represents the odds that a randomly selected student with features X_i attends a historically disadvantaged school. Using Bayes' rule we can simplify w_i to:

$$w_i = \frac{\Pr(X_i|g=0)}{\Pr(X_i|g=1)} = \frac{\Pr(g=0|X_i)/\Pr(g=0)}{\Pr(g=1|X_i)/\Pr(g=1)} \quad [3.5]$$

The estimates $\widehat{w}_i$ are easily computed using sample proportions of students in each group and predicted probabilities of group membership from a probability model for $\Pr(\mathrm{g} = 0|\mathrm{X_i})$. The decomposition is then performed using weighted regression estimates for the sample of historically advantaged schools, $\hat{\beta}_1^C$. In applying the reweighting approach to actual data with known sampling weights, w_i (where w_i is normalized to sum to 1), DiNardo et al (1996) propose using the product $\widehat{w}_i.w_i$ (also normalized so that the sum of the weights is equal to 1) as the weight in the counterfactual regression.

Through replacing the original counterfactual $\hat{\beta}_1$ with $\hat{\beta}_1^C$ we are able to precisely measure how much of the total performance gap can be explained by observable student and home background characteristics and how much of the gap is due to school resources and system functioning. In essence, this approach is similar to the inverse probability weighted (IPW) estimate of Hirano, Imbens, and Ridder (2003) commonly used in the program evaluation literature. The counterfactual of interest is therefore computed using students across the two school types who are truly comparable which allows for a result that more accurately relates to the PATN than the original two-fold OB decomposition.

Following Fortin et al (2011), the average performance gap can be represented by:

$$\hat{\beta}_1'\bar{X}_1 - \hat{\beta}_0'\bar{X}_0 = (\hat{\beta}_1'\bar{X}_1 - \hat{\beta}_1^{C\prime}\bar{X}_1^C) + (\hat{\beta}_1^{C\prime}\bar{X}_1^C - \hat{\beta}_0'\bar{X}_0) = \widehat{\Delta}_X + \widehat{\Delta}_S \quad [3.6]$$

The explained component $\widehat{\Delta}_X$ consists of two components:

$$\hat{\beta}_1'\bar{X}_1 - \hat{\beta}_1^{C\prime}\bar{X}_1^C = \hat{\beta}_1'\left(\bar{X}_1 - \bar{X}_1^C\right) + \bar{X}_1^{C\prime}(\hat{\beta}_1 - \hat{\beta}_1^C) \quad [3.7]$$

where the first term is a pure explained effect and the second part is due to specification error that results from assuming a linear model. Similarly, $\widehat{\Delta}_S$ consists of two components:

$$\hat{\beta}_1^{C\prime}\bar{X}_1^C - \hat{\beta}_0{}'\bar{X}_0 = \bar{X}_0{}'\left(\hat{\beta}_1^C - \hat{\beta}_0\right) + \hat{\beta}_1^{C\prime}(\bar{X}_1^C - \bar{X}_0) \quad [3.8]$$

where the first is a pure unexplained component and the second a reweighting error component that tends towards zero in large samples (Fortin et al, 2011). Given that the propensity scores are estimated only on student level information, we would not expect the reweighting error component of the unexplained gap to be zero. However, I believe that this representation of the unexplained gap provides a unique interpretation of the research question at hand and offers alternative interpretations to the components in [3.8].

The pure unexplained term $\bar{X}_0{}'\left(\hat{\beta}_1^C - \hat{\beta}_0\right)$ measures the expected performance difference due to differential functioning (returns structure) across the two school sub-systems, given the same level of educational inputs. For purposes of this study we will refer to this component as the school efficiency (SE) gap. The term $\hat{\beta}_1^{C\prime}(\bar{X}_1^C - \bar{X}_0)$ measures the expected performance difference due to dissimilar endowments of school level resources across the sub-systems, given the same level of functioning. For purposes of this study we will refer to this component as the school resources (SR) gap. We can think of the SE and SR gaps as relating to two separate "treatments" in the South African schooling system; one where the functioning of historically disadvantaged schools is augmented through interventions that are targeted at *inter alia* improving school management and institutional efficiency (SE gap); and another where school resources *inter alia* the quality of teachers and parental involvement are increased (SR gap).

Because of additive linearity in the potential outcome regression, it is fairly simple to compute the elements of the pure explained and SR components as detailed decompositions.[55] Denoting student and home background characteristics as H, school characteristics as S and teacher/classroom characteristics as T, we can rewrite the pure explained and SR gaps as:

$$\hat{\beta}_1{}'\left(\bar{X}_1 - \bar{X}_1^C\right) = \sum_{k=1}^{K}\hat{\beta}'_{1k}\left(\bar{H}_{1k} - \bar{H}_{1k}^C\right) + \sum_{l=1}^{L}\hat{\beta}'_{1l}\left(\bar{S}_{1l} - \bar{S}_{1l}^C\right) + \sum_{m=1}^{M}\hat{\beta}'_{1m}\left(\bar{T}_{1m} - T_{1m}^C\right) \quad [3.9]$$

$$\hat{\beta}_1^{C\prime}\left(\bar{X}_1^C - \bar{X}_0\right) = \sum_{k=1}^{K}\hat{\beta}_{1\mathrm{k}}^{C\prime}(\bar{H}_{1k}^C - \bar{H}_{0k}) + \sum_{l=1}^{L}\hat{\beta}_{1l}^{C\prime}\left(\bar{S}_{1l}^C - \bar{S}_{0l}\right) + \sum_{m=1}^{M}\hat{\beta}_{1\mathrm{m}}^{C}{}'(\bar{T}_{1m}^C - \bar{T}_{0m}) \quad [3.10]$$

where X comprises of K student and home background controls, L school controls and M teacher/classroom controls, respectively. Given that the propensity scores are estimated based on

[55] Interpreting a detailed decomposition of the unexplained component is however less straightforward than the explained component as issues arise when the explanatory variables of interest are categorical and do not have an absolute interpretation (Fortin, 2010). Tentative solutions which impose some normalisations on the coefficients have been supplied (see Gardeazabal & Ugidos, 2004; Yun, 2005, 2008), although interpretation may not be meaningful and further depends on the choice of reference group. Therefore, this study refrains from conducting detailed decompositions of the unexplained component.

student and home background characteristics, we would anticipate that $\bar{H}_{1k}^{C} \cong \bar{H}_{0k}$, $\bar{H}_{1k} \geq \bar{H}_{1k}^{C}$, $\bar{S}_{1l}^{C} \geq \bar{S}_{0l}$, $\bar{S}_{1l} \cong \bar{S}_{1l}^{C}$, $\bar{T}_{1m}^{C} \geq \bar{T}_{0m}$ and $\bar{T}_{1m} \cong T_{1m}^{C}$. The extent to which $\bar{S}_{1l} \cong \bar{S}_{1l}^{C}$ and $\bar{T}_{1m} \cong T_{1m}^{C}$ will depend on the distribution of students within the advantaged school system who possess characteristics similar to those of students within the disadvantaged school system. If they are predominantly attend lesser resourced schools (at least within a better resourced school sub-system), then after reweighting we would expect there to be some positive difference such that $\bar{S}_{1l} > \bar{S}_{1l}^{C}$ and $\bar{T}_{1m} > T_{1m}^{C}$. We can therefore think of the pure explained gap as reflecting the average performance difference that is due mainly to differences in student and home background characteristics.

Apart from matching, the methodological approach adopted by Botezat and Seiberlich (2013) differs from that of this paper in two important ways. First, it makes use of a threefold Oaxaca-Blinder decomposition that decomposes the gap into three parts as follows:

$$\bar{X}_1'\hat{\beta}_1 - \bar{X}_0'\hat{\beta}_0 = [E(Y^0|g=1) - E(Y^0|g=0)] + [E(Y^1|g=0) - E(Y^0|g=0)]$$
$$+[E(Y^1|g=1) - E(Y^1|g=0) - E(Y^0|g=1) + E(Y^0|g=0) \qquad [3.11]$$

where the counterfactual means $E(Y^0|g=1)$ and $E(Y^1|g=0)$ are estimated by the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964). The first term of [3.11] is the explained effect, the second term the unexplained effect and the final term captures the fact that the gap could be determined by the co-existence of different levels of individual characteristics and returns (interaction effect). The second important difference is that different counterfactuals are used to compute the explained and unexplained components of [3.11]. The unexplained components of [3.6] and [3.11] make use of the same counterfactual and are therefore comparable (barring choice of estimator). The explained components, however, are derived using different counterfactuals.[56] This part of the performance gap is not of as much interest as the relative size of the unexplained component as we should expect students in "wealthier" school systems to perform better given a higher likelihood of having come from more affluent households, and vice versa. One drawback of the three-fold decomposition as defined in [3.11] is that it does not allow a detailed decomposition of the explained component to be conducted.

One methodological question that might be posed is what is gained from including the interaction term. The author's conjecture would be that, dependent on the context under which

[56] The decompositions undertaken in this paper are framed from the point of view of the better performing school group (English/Afrikaans schools) whilst Botezat and Seiberlich (2013) frame their decompositions from the point of view of the worse performing school group.

achievement gaps are being assessed, the interaction effect may not be of interest. Rewriting the interaction term of [3.11] in terms of a reweighting estimator:

$$\left(\hat{\beta}_1'\bar{X}_1 - \hat{\beta}_0^C{}'\bar{X}_0^C\right) + \left(\hat{\beta}_0'\bar{X}_0 - \hat{\beta}_1^C{}'\bar{X}_1^C\right) \quad [3.12]$$

we would expect $\left(\hat{\beta}_1'\bar{X}_1 - \hat{\beta}_0^C{}'\bar{X}_0^C\right) > 0$ and $\left(\hat{\beta}_0'\bar{X}_0 - \hat{\beta}_1^C{}'\bar{X}_1^C\right) < 0$. The first component captures the expected performance difference between two students who attend schools within the different sub-systems but whose own and home background characteristics are similar to those of the average student attending a historically advantaged school. The difference in the expected performance of these two students results from differences in the distribution of school resources across the two school types as well as differences in the returns structure to school and student inputs. In the South African context we would expect both of these to be positive and potentially large. The interpretation is similar for the second component except here the comparison is between two students whose own and home background characteristics are comparable to that of the average student attending a historically disadvantaged school which is equivalent to the unexplained component in [3.11].

Essentially the interaction term captures the difference in the unexplained and explained effects in a two-fold OB decomposition that arise from using either one of the two counterfactuals $\hat{\beta}_0^C{}'\bar{X}_0^C$ and $\hat{\beta}_1^C{}'\bar{X}_1^C$. As a result, the sign of the net interaction effect will depend on how different the expected schooling environment (in terms of resources and returns structure) would be when moving an "average" student from their school system to that of their equal in another school system. In their estimations across eight European countries, Botezat and Seiberlich (2013) find interaction effects that are either small relative to the average gap or not significantly different from zero. In South Africa we would expect the net interaction term to be positive as the effect of moving an affluent student from the historically disadvantaged school sub-system into the better performing historically advantaged school sub-system would likely lead to a greater expected improvement in performance than would be observed from a similar movement of an impoverished student.

It should be mentioned that despite its relative simplicity, the reweighting method discussed above relies on a model specification for the propensity score that is adequately flexible in describing the relationship between pre-treatment characteristics and school attendance such that the approximation error is minimised. The most favoured estimation method is a parametric linear logistic regression with selected interactions and polynomial terms; variable choice is typically guided by economic theory, prior research and significance threshold "rules-of-thumb" (see Hirano & Imbens, 2001; Rosenbaum, 2002). I estimate $\Pr(X_i|g = 0)$ using a generalised additive model

(GAM) (Hastie et al., 2009) that replaces the linear link function in logistic regression with a flexible additive function. Propensity score estimation using GAM has been shown to lead to improved overall covariate balance when compared to logistic regression (see Woo, Reiter & Karr, 2008).[57]

3.3 Data and summary statistics

The Progress in International Reading Literacy Study (PIRLS) conducted in 2005/6 by the IEA[58] was the second of its kind conducted in a five year cycle (after PIRLS 2001) in which particular emphasis was placed on the reading proficiency of young children. Although the survey collected data on 45 schooling systems from 40 countries, only the South African data is used for purposes of this paper.[59] Grade 4 students were tested, with the exception of Luxembourg, New Zealand and South Africa, where students were sampled from the fifth grade. In addition to the collection of reading test scores, a full array of background information regarding home and school environments was gathered. The relatively large size of the South African dataset (14125 grade 5 students sampled from 385 schools) makes PIRLS 2006 highly advantageous for analysing educational outcomes and its determinants in South Africa, as previous research has revealed a very large intra-class correlation coefficient in South Africa of around 0.7 for reading scores (see for example Van der Berg, 2008). The sample of schools needs to be suitably large such that the sample variation in schooling outcomes truly reflects that observed in the South African education system. Of all the countries that participated in the PIRLS 2006 survey, the situation in South Africa proved to be the most complex given that the questionnaires and assessment tools had to be translated into all of the 11 official languages.

As this study is interested in the observed performance gap between historically black African and historically advantaged schools, the sample of students needed to be divided into these two school types. The dataset provides no information of the former school department, but schools were able to select the language of the test. It is safe to assume that schools that tested in an African language would have fallen under the historically black African system. It is furthermore likely that schools formerly belonging to the white, Indian and coloured education departments would have tested in English or Afrikaans. However, an overlap between the two groups may exist in that a

[57] This is particularly the case when the covariate distributions of the two groups have sufficient overlap. When sufficient overlap is lacking, GAM more clearly reveals this fact (Woo et al, 2008).
[58] International Association for the Evaluation of Educational Achievement.
[59] There may be concern that the developed country context of the PIRLS study may have generated a bias in the South African reading scores in favour of English speaking students in wealthier schools. However, similar performance gaps between rich and poor schools (as proxies for the former school departments) have been observed in regional studies (c.f. van der Berg, 2008; Spaull, 2013).

number of formerly black African schools may have tested in (particularly) English.[60] Therefore I will refrain from using the distinction of former disadvantaged and advantaged schools and rather denote the groups as English/Afrikaans testing schools and African language testing schools. In order to address the issue of overlap between the two groups, a further restriction was applied to the sample of formerly advantaged schools. If more than 65 percent of the grade 5 sample from a particular school was found to not speak the test language on a regular basis, this school was dropped from the group of English/Afrikaans testing schools.

The decision to drop schools and not simply move them to the sample of African language testing schools was made as some of the schools meeting the aforementioned restriction may not in fact be historically black African schools. In fact, some of the schools may be historically coloured schools that are poor and weak performing. Consequently, the remaining sample of English/Afrikaans testing schools may suffer from positive selection bias if we assume that the remaining group of schools are the richer, and hence better performing schools. This should be kept in mind when interpreting the results. Estimates based on the full sample of English/Afrikaans testing schools will serve as a robustness check to the main results.

The dependent variable employed in the empirical model is the individual student reading score.[61] The main problem posed by the data was that of a large number of missing data, particularly at the student level. Dropping these students would reduce the amount of variation in the dependent variable, causing bias in the results (Ammermueller, 2006). A brief note on the imputation methods used to deal with missing data at the household level is provided in note 3.1 of the appendix to this chapter. Given the comparatively smaller number of missing data at the school levels, schools with missing data were dropped from the sample. Definitions of all controls variables included in the empirical model are provided in table A3.1 of the appendix. The final sample includes 9134 students in 240 African language testing schools, and 2107 students in 66 English/Afrikaans testing schools. This is similar to what is observed in the South African education system: 21 percent formerly “advantaged” schools and 79 percent formerly black African (disadvantaged) schools.[62]

[60] In a separate study by (Desai, 2001), a primary school in the Khayelitsha township, Cape Town, was observed where the home language of the majority of students and educators was Xhosa. However, since 1995 the school has decided to use English as the medium in which all school work is to be expressed from grade 4, although this does not prevent the teachers from relaying information to the students in an African language.

[61] The test score is calculated using average scale scores computed from 5 plausible imputed scores based on Item Response Theory (IRT). The international scores are set on a scale with a mean of 500 and a standard deviation of 100.

[62] “Advantaged” here refers to schools that did not fall under the former black African (DET and homeland) school system and therefore may include former white, coloured and indian schools. The former DET and homeland schools make up approximately 80 percent of South African primary schools. Some of the “advantaged” schools, particularly coloured schools, are not likely to be wealthy, well-functioning schools.

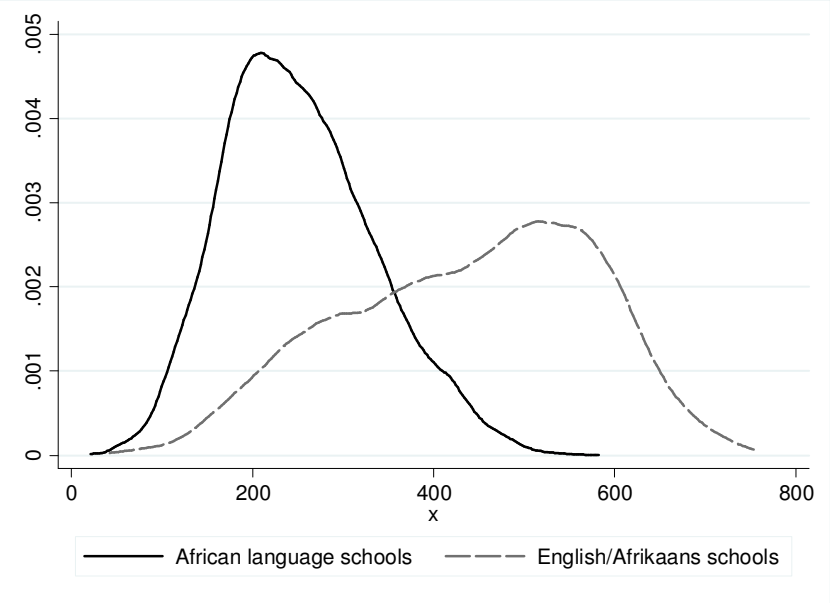
The sample average reading scores are 252 and 465 for African language and English/Afrikaans testing schools respectively, representing a statistically significant performance gap of 213 points. This difference is dramatic when viewed in the context of 50 points on the PIRLS test being described as equivalent to one school grade (Filmer, Hasan and Pritchett, 2006). South African students attending both school types performed lower than the international average and both a higher mean and test score spread for English/Afrikaans testing schools is depicted in Figure 3.1. Figure A3.1 of the appendix shows the distribution of African language testing schools compared to English/Afrikaans testing schools with and without sample restrictions. It is clear that the excluded schools are predominantly worse performing ones, yet even after the restrictions are made a significant proportion of students in the group of English/Afrikaans testing schools are performing at quite low levels.[63] This is due to the fact that the former may include a number of coloured schools that may have similarly low SES levels as African language testing schools. This is particularly the case among coloured schools that are comparatively poorer than their affluent counterparts within the same school grouping.

Standardised differences in means shown in Figure A3.5 (black bars) indicate clear differences in the composition of the student body and allocation of school resources across the two school types. Specifically, comparisons of the means indicate that students attending English/Afrikaans testing schools are significantly less likely to be overage as well as significantly more likely to speak the test language at home on a regular basis. Furthermore, students attending English/Afrikaans testing schools are more likely to receive help with their reading homework, have better educated parents with full-time employment and come from households with higher socio-economic status (SES)[64]. African language testing schools report higher levels of absenteeism and are significantly poorer on average as measured by the average SES of the student body. Surprisingly, a significantly larger proportion of teachers in African language testing schools report a greater variety of daily use of in-class learning and teaching activities and methods and diagnostic testing. However, this does not allude to how much time is spent on each activity (or even the quality of the activity), which might vary between schools.

[63] Figure A3.2 compares the reading score performance of the two school groups under consideration in this study to the literacy test score distributions of grade 4 students by former department from the NSES study conducted in 2008. Typically former department is proxied by school wealth (see for example Van der Berg, 2008; Taylor & Yu, 2008; Spaull, 2013). However, from figure A3.2 it appears that the division based on language of test proxies closer to the former white school system than using the top 20 percent wealthiest schools based on average school SES.

[64] The socio-economic status of a student's household is measured using first principal component analysis of 10 household assets.

Figure 3.1: Reading score distributions, by school type



Notes: own calculations using PIRLS (2006)

As mentioned, the propensity score model includes student and household background characteristics (as listed in table A3.1) as controls. The standard deviation of student household SES within a school is also included as a control. From figure A3.3 it is clear that there may be some overlap issues across the two school groups with regards to the average SES of students within schools (school SES), although the same cannot be said of the standard deviation. Furthermore, the standard deviation in household SES is narrower in the extremes of school average SES; that is, the wealthiest and poorest schools are characterised by relatively equal distribution of household SES across students. It is the belief of the author that the inclusion of the standard deviation of student household SES in the propensity score model leads to better matches, particularly when focus is placed on the most comparable students and schools across the two subsystems who are less likely to fall in the extremes of school wealth. The robustness of the results for the inclusion of this variable in the propensity score model will be tested. The grey bars in figure A3.5 of the appendix illustrates that inverse probability weighting of the group of English/Afrikaans students provides a reweighted treatment sample that is much more closely balanced with the control group with regards to student and home background factors. As expected, there remain substantial differences in school and teacher characteristics.

The propensity score distributions (histograms) of the two school types presented in figure A3.4 of the appendix suggest acceptable common support, although insufficient overlap in the extremes of $Pr(X_i)$ may need to be addressed. If particular values of $Pr(g = 0|X_i)$ are rare among

$g = 1$ and common among $g = 0$, such observations will receive a very large weight in the estimator. This has implications for estimation bias of the counterfactual and treatment effect (Dehejia & Wahba, 1999; Heckman, Ichimura & Todd, 1998). Nõpo (2008) points out that should there be a lack of common support in the covariates, the unexplained component as an estimate of the treatment effect will be upward biased. To address issues of overlap, I test the robustness of my results using a simple selection rule that excludes observations whose propensity scores fall outside of the [0.1, 0.9] range (Crump, Hotz, Imbens & Mitnik, 2009). Results from kernel matching and nearest neighbour matching procedures are also estimated for comparison purposes. All estimate standard errors are obtained using 500 bootstrap iterations.

3.4 Empirical results

3.4.1 Aggregate decomposition results

Estimates of the explained and unexplained components shown in column 1 of table 3.1 (panel b) indicate that 81.2 percent of the test score gap between African language testing and Afrikaans/English testing schools can be explained by differences in average endowments of student, household and school characteristics. The remaining 18.8 percent represents the unexplained gap that is due to differences in school efficiency. Both the explained and unexplained gaps are statistically significant. The OB decomposition results therefore suggest that former advantaged schools and their students are both more endowed with characteristics conducive to higher schooling outcomes and more efficient in transforming educational inputs into educational outcomes. In other words, keeping the distribution of characteristics of African language schools and their students the same but facing the English/Afrikaans school returns to these characteristics, students' test scores would be improved by an average of 40 points.

As can be seen from equation [3.3], the size of the explained component in the OB decomposition is dependent on two factors: the difference in the average endowments between the two school types and the coefficient structure of the English/Afrikaans testing schools. Endogeneity biases due to non-random selection may bias the latter, most likely in an upwards direction. For example, exclusion of predominantly weaker historically advantaged schools from the sample of English/Afrikaans testing schools and selection into the wealthier and conceivably better performing former advantaged school system driven by unobservable factors which are positively related to schooling inputs and processes. Given the large positive coefficients on, in particular, school SES, parent involvement and teacher education in the English/Afrikaans schools model,[65] as well as the

[65] These regression results are not shown here but are available from the author on request.

higher average endowments in favour of these schools, it is unsurprising that the decomposition yields a large and significant explained component. Correction for selection and endogeneity biases may result in different relative sizes of the explained and unexplained coefficients.

Table 3.1: Aggregate decomposition results

Panel a:	Average test score		Observations	
English/Afrikaans schools	464.6		2107	
African language schools	251.5		9134	
Average test score gap	213.1			
Panel b:	(1) OB decomposition		(2) Reweighted decomposition	
Decomposition components:	Estimate	Proportion of gap	Estimate	Proportion of gap
Explained gap	173.0*** (10.97)	81.2	123.2*** (5.23)	57.8
Pure explained gap			122.5*** (5.47)	57.5
Specification gap			0.7 (1.69)	0.3
Unexplained gap	40.0*** (10.95)	18.8	89.8*** (5.80)	42.2
School efficiency gap			30.2** (14.18)	14.2
School resource gap			59.6*** (14.60)	28.0

Notes: Bootstrapped standard errors computed from 500 repetitions shown in parentheses. OB decomposition components are estimated using the English/Afrikaans school returns as counterfactual. Reweighted decomposition components are estimated using the returns structure from the reweighted English/Afrikaans sample as counterfactual. *p<0.10, **p<0.05 and ***p<0.01.

Alternatively, the explained component is biased by the inclusion of post-treatment variables which are related to the effect of treatment and not selection into treatment. A reweighted decomposition that corrects for pre-treatment differences, as well as decomposes the unexplained effect into two parts (one due to school resource differences and another due to school efficiency differences) will provide a more accurate reflection of the average treatment of attending a historically advantaged school. The results of a reweighted decomposition are shown in column 2 of table 3.1 (panel b).

Following reweighting, the pure explained gap that captures the achievement differential due to (mainly) differences in student home background is estimated to be 122.5 points, or 57.5 percent of the average performance gap. This estimate is substantially lower than the explained component of the OB decomposition which accounted for 81.2 percent of the average performance gap. The statistically insignificant estimate of the specification error accounts for only 0.3 percent of

the overall gap, suggesting a truly linear test score model. The total unexplained gap now accounts for 42.2 percent of the average test score gap as opposed to 18.8 percent in the original decomposition model. This was anticipated given that part of the "treatment" of attending an English/Afrikaans school is attributed to higher endowments of school resources contained within the explained component of the original decomposition.[66] This aspect of treatment (termed SR gap) makes up 28 percent of the average test score gap, whilst school efficiency differences contribute 14.2 percent.

Table 3.2 summarises the detailed decomposition results of the pure explained and SR gaps. Average school SES and province are separated from other school characteristics. It is the opinion of the author that unlike the teacher and classroom level controls listed in table A3.1, the majority of the school resources controlled for in the outcome model are (to different degrees) under the control of government and therefore policy. For example, parent involvement and absenteeism are related to the efficacy of school management and accountability which may in part be attributed to training and hiring practices. Another example is school SES which is believed to capture aspects of school resourcing that may not be related to public funding such as smaller class sizes, as well as serve as a proxy for institutional and cultural processes related to effective school management and governance.

As expected, student characteristics play no role in determining the size of the SR gap but contribute to a significant proportion of the pure explained gap, close to 30 percent of the total average performance gap. A further important contributing factor to the explained gap is average school SES. The results of table 3.2 suggest that if we were to compare two students within the English/Afrikaans school group, where one student is comparable to the average student found within the group of African language schools, then differences in the home background and affluence of the immediate peer group of these two students would account for as much as half of the expected performance gap between these two students. This result is unsurprising as there are many mechanisms in place that prevent poor children from attending the most affluent schools.[67]

The results are also in agreement with the documented "flight" of more affluent black African students out of historically black schools, with little if any movement in the opposite direction (Chisholm, 2004: 104).[68] Consequently, black schools are left with the poorest members of

[66] Note that the combined contribution of the explained gap and the school resource gap of 85.5 is very similar to the total explained gap of the OB decomposition in table 3.1.
[67] For example, many of the affluent schools in South Africa charge fees to cover the costs of schooling not borne by the state. This power to charge fees creates an incentive to admit as many full fee-paying students as the school can accommodate (Woolman & Fleisch, 2006).
[68] An example of this is provided in an article by Woolman and Fleisch (2006). They describe how Sandown High in Sandton, Gauteng, is oversubscribed whereas on the other side of town in Orlando High, Soweto, classrooms stand

the community (Chisholm, 2004: 106). This is partly reflected through the contribution of peer socio-economic status to the SR gap where 18.5 percent of the average gap is estimated to stem from peer affluence differences between comparable students (in terms of own and home background characteristics) across the two school groups. The results therefore suggest that social factors - such as the socio-economic status of the peers that South African primary school students find themselves in class with - play a considerable role in determining achievement, and that segregation along socio-economic lines explains a substantial proportion of average performance differentials in the South African school system.

The small and insignificant contribution of school resources to the expected explained gap indicates that after reweighting on the propensity of attending an African language school, school resources are equivalently distributed (at least on average) across students attending English/Afrikaans schools. It is not unsurprising that differences in class/teacher resources contribute positively to the explained gap. Students attending English/Afrikaans schools who are comparable in own and home characteristics to students attending African language schools are more likely to attend schools where school governing body[69] funding of, for example, teachers and classrooms is less likely to be augmented through higher school fees.

The insignificant contribution of teacher and classroom factors to the SR gap suggests equal average endowments of these factors across former departments. Differences in the distribution of school resources that may be indicators of school leadership and management account and school SES account for 60 points of the SR gap, or 28 percent of the total expected performance gap. The results therefore suggest that should both the processes that encourage better functioning and the effectiveness of former disadvantaged schools be improved to the level of former advantaged schools, the average performance gap might be closed by as much as 90 points (42 percent).

3.4.2 Sensitivity checks

Given concerns of insufficient overlap, the reweighted decomposition is re-estimated considering only those students whose estimated propensity scores fall within the range [0.1, 0.9] (Crump et al, 2002). This trimming procedure reduces the samples of g = 0 and g = 1 students under consideration by approximately 55 and 20 percent respectively. Over this range, the unexplained gap contributes towards 65.5 percent of the average test score gap with 15.9 percentage points attributable to differences in school efficiency (see column 1 of table 3.3). The largest proportion of the average gap

empty. Many of the students attending Sandown High reside close to Soweto in the Alexandra township, yet they choose to travel many kilometres to attend school elsewhere.

[69] Parent-elected school governing bodies (SBGs) are conferred authority through the South African Schools Act (SASA) in matters such as admissions policy, school fees and staff appointments.

is therefore due to differences in school resources, although only a third of the SR gap is explained by differences in factors that could be directly influenced by policy (see table 3.4). Overall, approximately a third of the average test score gap between the two school systems might be closed by bringing the school efficiency and school resources of African language schools in line with those of English/Afrikaans schools.

Table 3.2: Detailed aggregate decomposition results

	(1)		(2)			
	OB decomposition		**Reweighted decomposition**			
Characteristics	Explained gap	Proportion of total gap	"Pure" explained gap	Proportion of total gap	School resource gap	Proportion of total gap
Student/house-hold	59.00***	27.7	60.9**	28.6	-2.4	-1.1
	(4.03)		(4.30)		(2.99)	
School	14.1**	6.6	1.9	0.9	20.2***	9.5
	(6.39)		(2.53)		(6.88)	
School SES	74.3***	34.9	45.0***	21.1	39.4**	18.5
	(11.65)		(3.89)		(13.65)	
Class/teacher	12.9***	6.1	7.4***	3.5	2.1	1.0
	(3.38)		(2.11)		(4.35)	
Province	12.7***	6.0	7.2***	3.4	10.2*	4.8
	(4.01)		(2.26)		(5.56)	
Total	173.0***	81.2	122.5***	57.5	59.6***	28.0
	(10.97)		(5.47)		(14.60)	
Observations (g = 0)	9134		9134			
Observations (g = 1)	2107		2107			

Notes: Bootstrapped standard errors computed from 500 repetitions shown in parentheses. OB decomposition components are estimated using the English/Afrikaans school returns as counterfactual. Reweighted decomposition components are estimated using the reweighted English/Afrikaans school returns as counterfactual. * p<0.10, **p<0.05 and ***p<0.01.

Further sensitivity checks allow for alternative propensity and outcome model specifications. In all further specifications only the sample where the estimated propensity score falls in the range [0.1, 0.9] is used for estimation.[70] Excluding the standard deviation of SES within schools from the propensity score model transfers about 3 percent from the explained gap to the school efficiency gap (see column 2 of table 3.3). Aside from this, the detailed decomposition results are robust to the original model (see table 3.4). Excluding province from the outcome model does not alter the explained gap but results in a transfer from the SE gap to the SR gap within the explained gap.

[70] Sensitivity checks where estimation uses the full sample provide results that are generally robust to those provided in tables 1 and 2. These are available from the author on request.

Furthermore, a transfer within the SR gap occurs away from school resources and towards school SES and teacher/classroom factors. These changes cannot be driven by differences in average endowments of S (including school SES) and T as the exclusion of provincial dummies in the outcome model in no way alters these. Rather, the changes occur through $\hat{\beta}_1^C$ suggesting that certain school and teacher resources as well as student SES are not randomly distributed along provincial lines. Exclusion of province from the outcome model will result in bias in the returns to these inputs.

Correction for sampling weights also has implications for the relative sizes of the school efficiency and school resources gaps. From table 3.3 (column 4) we can see that whilst the proportions of the total gap ascribed to the explained and unexplained components are fairly robust, the SE gap increases and the SR gap decreases when sampling weights are ignored. One reason for this may be due to the number of schools sampled from each province; for example, former advantaged schools within the Northern Cape Province make up approximately 9 percent of all former advantaged schools (Department of Basic Education, 2013) yet Northern Cape schools sampled in the PIRLS 2006 dataset make up 19 percent of all English/Afrikaans testing schools surveyed. This will have implications for the estimated model coefficients as the Northern Cape tends to be one of the weaker performing provinces. From table 3.4 it appears that the transfer from the SR to the SE gap works primarily through changes in the contribution of school SES to the former. It is therefore important to consider the role of sampling when interest lies in extracting detailed information about the unexplained component of the test score gap.

Columns 5 and 6 of table 3.3 indicate the results from kernel matching and nearest neighbour matching procedures in order to check for upward bias in the unexplained effect. Results from the reweighted decomposition without sampling weights are used for comparison as there is no clear method for accommodating sample weights in the matching literature. For consistency, only observations with $0.1 < p(x) < 0.9$ are considered. Busso et al (2013) have shown that nearest neighbour matching generally has small bias. In comparing the nearest neighbour to the reweighted decomposition, the estimates appear to suggest that there may be a degree of upward bias in the reweighted decomposition estimate of the unexplained effect. However, the reweighted decomposition results do not differ significantly from that of kernel matching.

Table 3.3: Sensitivity checks based on propensity score selection rules, alternative model specifications and matching procedures

Type of decomposition	(1) Reweighted		(2) Reweighted (standard deviation of SES excluded from propensity model)		(3) Reweighted (province dummies excluded from outcome model)	
Common support restriction?	0.1 < p(x) < 0.9		0.1 < p(x) < 0.9		0.1 < p(x) < 0.9	
Average test score gap	201.1		205.7		201.1	
Decomposition components:	Estimate	% of gap	Estimate	% of gap	Estimate	% of gap
Explained gap	69.5*** (3.42)	34.5	64.1*** (3.21)	31.2	69.5*** (3.42)	34.5
Pure explained gap	68.2*** (3.34)	33.9	63.1*** (3.08)	30.7	67.7*** (3.34)	33.7
Specification gap	1.3 (0.92)	0.6	1.0 (0.90)	0.5	1.8 (0.92)	0.8
Unexplained gap	131.7*** (4.59)	65.5	141.6*** (4.71)	68.9	131.7*** (4.59)	65.5
School efficiency gap	32.0** (13.42)	15.9	44.3*** (13.79)	21.6	24.7** (10.90)	12.3
School resource gap	99.7*** (13.15)	49.6	97.3*** (13.58)	47.3	107.0*** (11.02)	53.2
Observations g = 1	1745		1757		1745	
Observations g = 0	3983		4119		3983	

Notes: Bootstrapped standard errors computed from 500 repetitions shown in parentheses. Reweighted decomposition components are estimated using the reweighted English/Afrikaans school returns as counterfactual. Kernel matching makes use of Silverman's rule-of-thumb for bandwidth selection. * p<0.10, **p<0.05 and ***p<0.01.

Table 3.3 continued: Sensitivity checks based on propensity score selection rules, alternative model specifications and matching procedures

Type of decomposition	(4) Reweighted (no sample weights)		(5) Kernel matching		(6) Nearest neighbour matching (k = 3)	
Common support restriction?	0.1 < p(x) < 0.9		0.1 < p(x) < 0.9		0.1 < p(x) < 0.9	
Average test score gap	159.1		159.1		159.1	
Decomposition components:	Estimate	% of gap	Estimate	% of gap	Estimate	% of gap
Explained gap	59.8*** (2.6)	37.6	62.7*** (2.72)	39.4	65.7*** (2.94)	41.3
Pure explained gap	59.2*** (2.51)	37.2				
Specification gap	0.60 (0.77)	0.4				
Unexplained gap	99.4*** (3.84)	62.4	96.4*** (3.87)	60.6	93.4*** (1.57)	58.7
School efficiency gap	41.9*** (8.30)	26.3				
School resource gap	57.4*** (8.06)	36.1				
Observations g = 1	1745		1745		1745	
Observations g = 0	3983		3983		3983	

Notes: Bootstrapped standard errors computed from 500 repetitions shown in parentheses. Reweighted decomposition components are estimated using the reweighted English/Afrikaans school returns as counterfactual. Kernel matching makes use of Silverman's rule-of-thumb for bandwidth selection. * p<0.10, **p<0.05 and ***p<0.01.

Table 3.4: Detailed decomposition results for different model specifications

	(1) Observations		(2) Student/ household		(3) School	
	g = 0	g = 1	Estimate	% of gap	Estimate	% of gap
Pure explained gap						
Reweighted decomposition	3983	1745	31.7*** (2.73)	15.8	0.40 (1.75)	0.2
Other specifications:						
Standard deviation of SES excluded from propensity model	4119	1745	31.2*** (2.76)	15.2	-1.40 (1.75)	-0.7
Province excluded from outcome model	3983	1745	31.9*** (2.75)	15.9	2.00 (1.61)	1.0
No sampling weights	3983	1745	29.1*** (2.10)	18.3	1.30 (1.35)	0.8
School resource gap						
Reweighted decomposition	3983	1745	2.40 (2.24)	1.2	28.0*** (8.45)	13.9
Other specifications:						
Standard deviation of SES excluded from propensity model	4119	1745	4.70* (2.41)	2.3	25.6*** (8.89)	12.5
Province excluded from outcome model	3983	1745	2.30 (2.30)	1.1	15.3*** (5.62)	7.6
No sampling weights	3983	1745	-1.00 (1.65)	-0.6	18.8*** (6.28)	11.8

Notes: Bootstrapped standard errors computed from 500 repetitions shown in parentheses. Reweighted decomposition components are estimated using the returns structure from the reweighted English/Afrikaans sample as counterfactual. All specifications are computed over the estimated propensity score range [0.1, 0.9]. *p<0.10, **p<0.05 and ***p<0.01.

Table 3.4 continued: Detailed decomposition results for different model specifications

	(4) School SES		(5) Teacher/classroom		(6) Province	
	Estimate	% of gap	Estimate	% of gap	Estimate	% of gap
Pure explained gap						
Reweighted decomposition	29.2*** (2.87)	14.5	5.60*** (1.78)	2.8	1.30 (1.19)	0.6
Other specifications:						
Standard deviation of SES excluded from propensity model	28.3*** (2.74)	13.8	4.8*** (1.79)	2.3	0.20 (1.12)	0.1
Province excluded from outcome model	27.9*** (7.75)	13.9	6.00*** (1.70)	3.0	-	-
No sampling weights	22.8*** (1.95)	14.3	5.30*** (1.63)	3.3	2.20** (0.95)	1.4
School resource gap						
Reweighted decomposition	42.7*** (12.16)	21.2	3.60 (4.64)	1.8	23.0*** (6.41)	11.4
Other specifications:						
Standard deviation of SES excluded from propensity model	38.7*** (12.42)	18.8	7.00 (4.82)	3.4	21.4*** (6.27)	10.4
Province excluded from outcome model	76.6*** (9.76)	38.1	12.8*** (4.39)	6.4	-	-
No sampling weights	20.6** (8.21)	12.9	3.90 (4.47)	2.5	15.6*** (5.54)	9.8

Notes: Bootstrapped standard errors computed from 500 repetitions shown in parentheses. Reweighted decomposition components are estimated using the returns structure from the reweighted English/Afrikaans sample as counterfactual. All specifications are computed over the estimated propensity score range [0.1, 0.9]. *p<0.10, **p<0.05 and ***p<0.01.

3.5 Concluding remarks

This study aimed to analyse the PIRLS reading score gap between the students of former black African and former advantaged schools. A semiparametric procedure that relaxes the functional form assumptions of the OB decomposition was employed. The methodological contribution of this paper was to separate the average test score gap into three parts: one due to differences in student and household characteristics (explained gap), another due to differences in the distribution of school resources (school resource gap) and another that is due to differences in the returns structure across school types (school efficiency gap). The use of reweighting to construct the counterfactual means further allowed detailed decompositions to be conducted on the explained and school resource gaps.

Comparison of the reweighted regression procedure of DiNardo (2002) to the traditional OB decomposition illustrates that use of the latter is likely to overstate the explained component of the performance gap. The estimated explained effect of 81.2 percent under the OB decomposition dramatically outweighs that of 57.8 percent estimated with reweighting. Issues of overlap were also able to be addressed as reweighting on propensity scores combined with trimming allows us to constrain the sample to believably more comparable groups of students across the two sub-systems. Trimming the sample to include observations with propensity scores falling within a [0.1, 0.9] range resulted in a smaller explained component of 34.5 percent of the average reading gap, with the remaining 65.5 percent considered to be representative of the treatment of attending a historically advantaged school. All policy relevant conclusions that follow consider only the estimates based on the trimmed sample. Whilst this may not provide results that are generalizable to the entire schooling system, it does provide interesting insights into the "treatment" of attending a historically advantaged school.

Home background factors are estimated to play a significant role in explaining the test score gap. However, of more relevance to policy is the role of school resources and school functioning as inequality in the home backgrounds of students will likely take generations to address. Between 14 and 35 percent of the expected test score gap is accounted for by the contribution of differences in school level controls to the school resource gap, depending on whether or not school SES is included. The results also provide evidence that, at least on average and for comparable students, the distribution of observable teacher and classroom factors controlled for in this study are fairly balanced across former departments. This is not to say that the quality of teaching and classroom processes are equal across sub-systems. Quality differentials that are captured by the school efficiency gap is estimated to account for 16 percent (32 points) of the average performance gap.

Overall, the decomposition results estimated here predict that successfully addressing inequalities in the distribution of school resources (or processes) that augment performance whilst simultaneously addressing inequalities in school effectiveness or quality may as much as halve the average performance gap between the two former school departments.

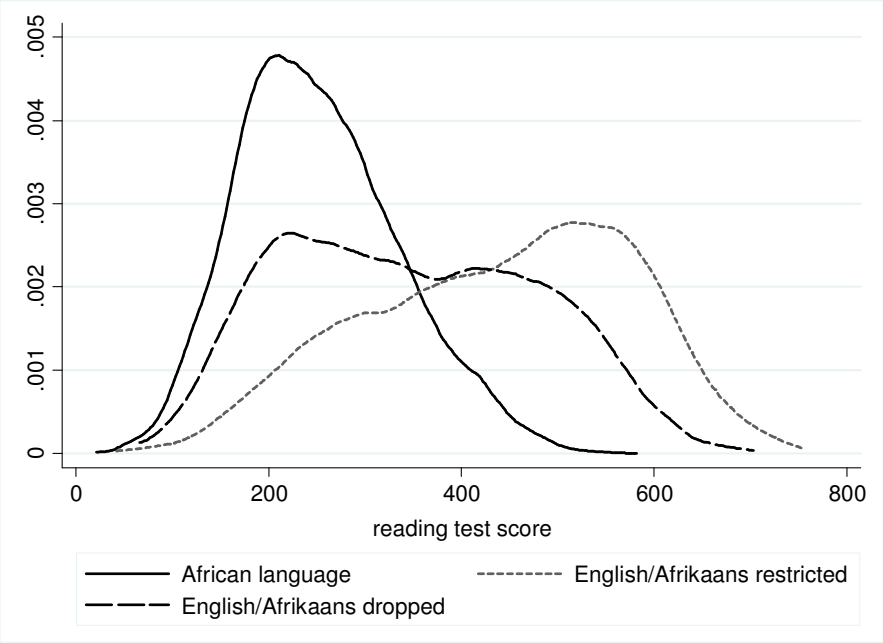
Policy targeted at improving schooling outcomes may prove ineffectual if the simultaneity mentioned above is ignored. Improving the efficiency of schools may be contingent on the necessary institutional and managerial processes already being in place. Recent analysis by Taylor (2011) of South African primary school outcomes using the National School Effectiveness Study (NSES) found that whilst school resource variables tend to be insignificant determinants of achievement, indicators of school management were consistently related to test scores. This suggests that the impact of school resources may be conditional upon how well those resources are managed (Van der Berg, 2008). The specific indicators of effective management controlled for in this study should not be interpreted as more than indicators that point to the characteristics typically exhibited by good managers, rather than levers to be manipulated by policy to achieve improved outcomes. For example, encouraging effective parent involvement through membership of a school governing body relies on equitable balance of power between educated staff on the one hand and, in many cases, illiterate parents lacking in the capacity to contribute to decision making processes on the other. A better route for policy would be to explore ways to attract and train better teachers and principals, as well as to cultivate an environment whereby accountability and the encouragement and empowerment of better teaching and school leadership can succeed.

Appendix to Chapter 3

Note 3.1

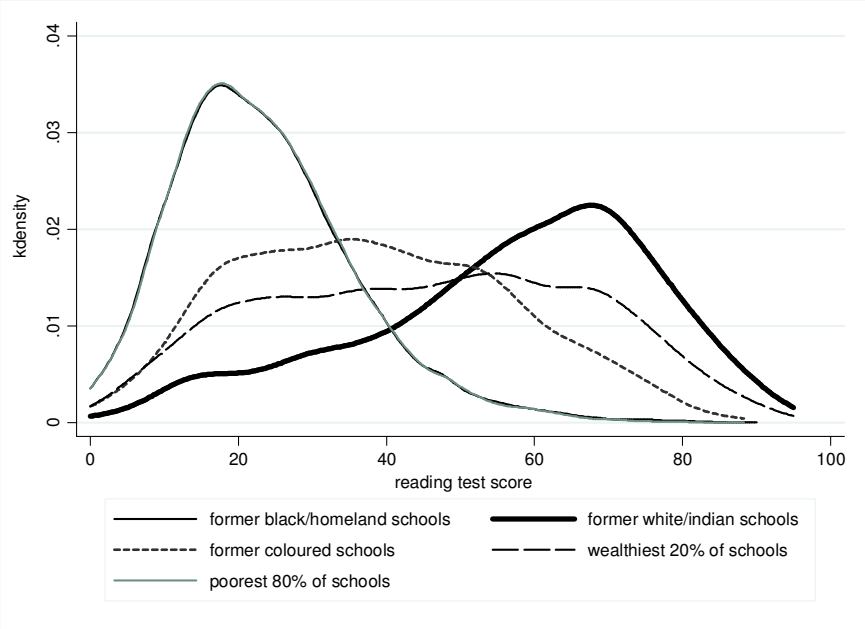
1) Missing data on possession items: missing values on household asset ownership were imputed using average possession within each of the 62 explicit strata (according to province and language). Household SES was subsequently estimated using first principal component analysis (PCA) and then standardised to have a mean of zero and a standard deviation of 1. School SES was calculated as the mean household SES within school and also standardised to have a mean of zero and a standard deviation of 1.
2) Missing data on other student and household characteristics: "missing/unspecified" was grouped as a separate category and a dummy variable coded "1 = missing/unspecified, 0 = otherwise" was included as a control in the regression model. In most cases, the coefficients on these "missing/unspecified" dummy variables were not found to be significantly different from the reference category. Missing data on categorical variables were therefore grouped with the reference category.
3) Missing values on parent education: imputed using the median parental education of the school.

Figure A3.1: Distributions of reading scores (weighted) by school type



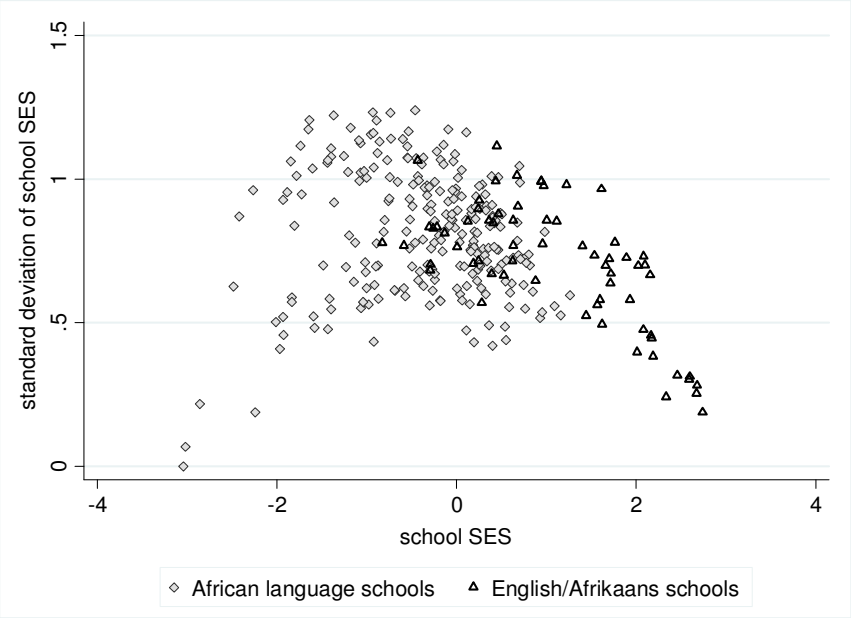
Notes: own calculations using PIRLS grade 4 reading scores (2006)

Figure A3.2: Distributions of reading scores (weighted) by former department



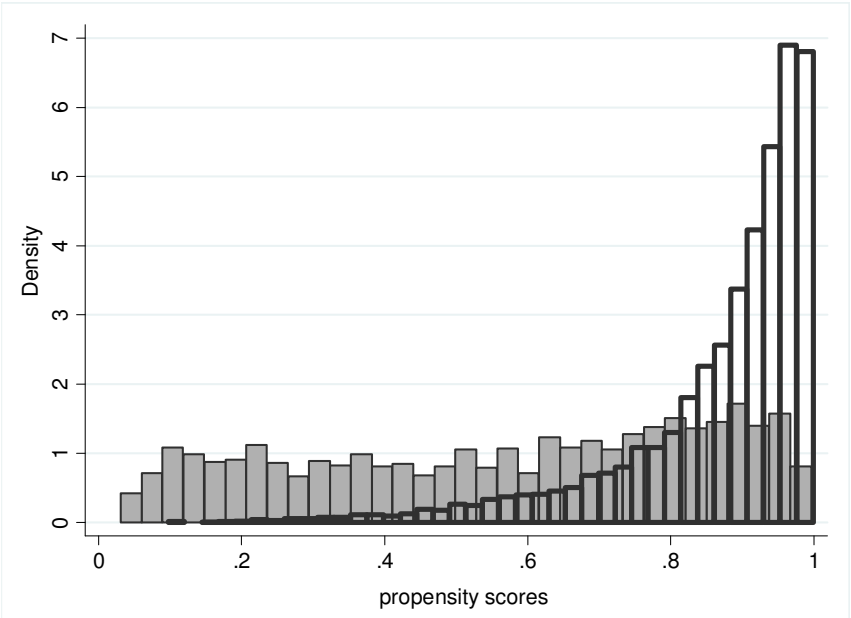
Notes: own calculations using NSES grade 4 reading scores (2008)

Figure A3.3: Distribution of SES by school type



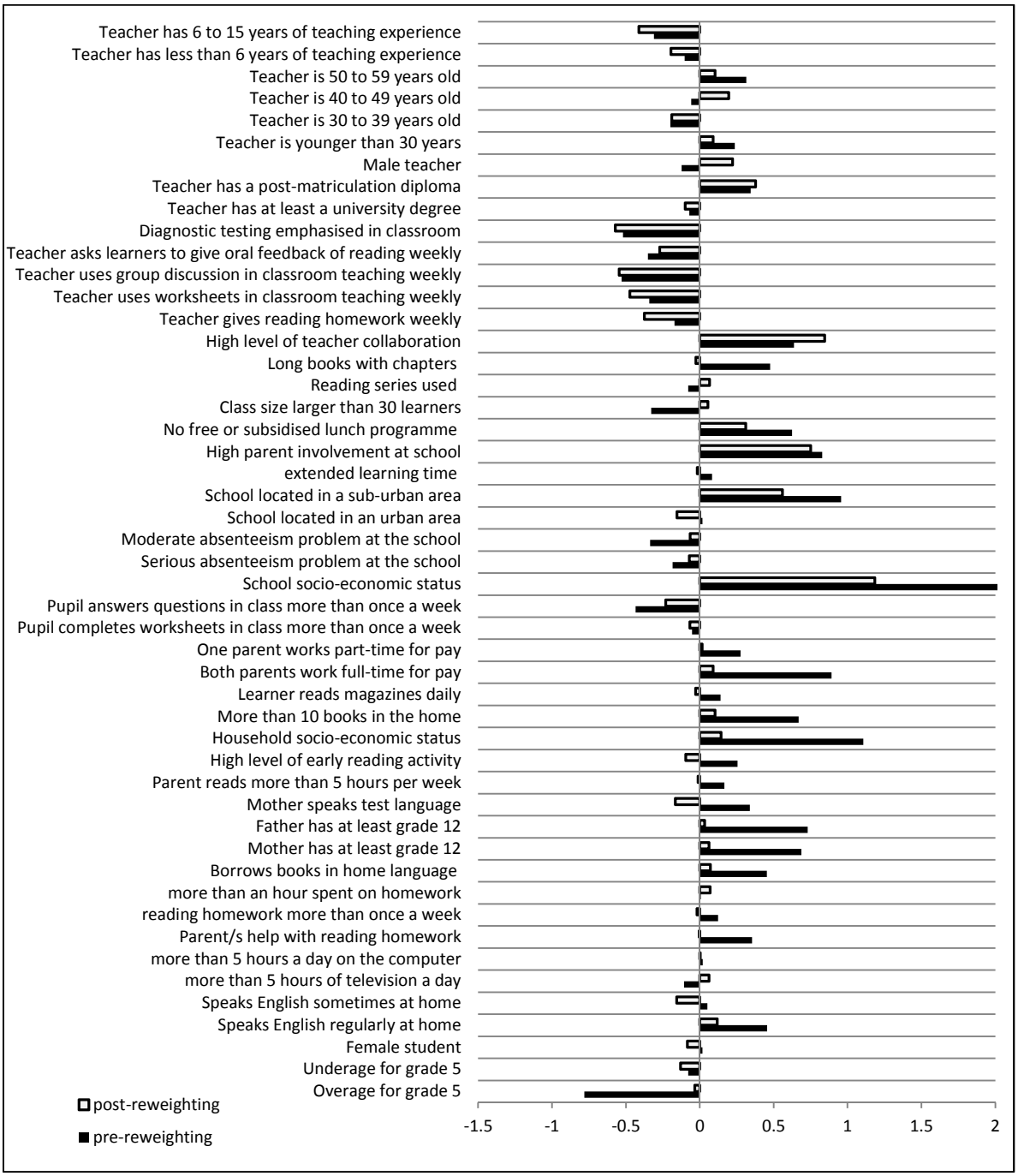
Notes: own calculations using PIRLS grade 4 reading scores (2006)

Figure A3.4: Estimated propensity scores by school type



Notes: own calculations using PIRLS grade 4 reading scores (2006); Grey bars represent English/Afrikaans schools, clear outline bars represent African language schools.

Figure A3.5: Standardised difference in means pre- and post-reweighting



Notes: covariate balance represented as standardised differences in means across control (African language schools) and treatment (English/Afrikaans schools) students for each of two samples.

Table A3.1: Descriptions of control variables

Variable	Description
Overage for grade 5	Dummy (0,1)
Underage for grade 5	Dummy (0,1)
Student is female	Dummy (0,1)
Student speaks English regularly at home	Dummy (0,1)
Student speaks English sometimes at home	Dummy (0,1)
Watches more than 5 hours of television a day	Dummy (0,1)
Spends more than 5 hours a day playing games on the computer	Dummy (0,1)
Parent/s help with reading homework	Dummy (0,1)
Receive reading homework more than once a week	Dummy (0,1)
Spends more than an hour on reading homework	Dummy (0,1)
Borrows books in home language outside of school	Dummy (0,1)
Mother has at least a matriculation qualification	Dummy (0,1)
Father has at least a matriculation qualification	Dummy (0,1)
Mother speaks the test language at home	Dummy (0,1)
Parent/s read for more than 5 hours per week at home	Dummy (0,1)
High level of early reading activity [a]	Dummy (0,1)
Household socio-economic status	Continuous (mean = 0, s.d. = 1)
More than 10 books in the household	Dummy (0,1)
Student reads magazines on a daily basis	Dummy (0,1)
Both parents work full-time for pay	Dummy (0,1)
One parent works part-time for pay	Dummy (0,1)
Teacher is male	Dummy (0,1)
Teacher is younger than 30 years	Dummy (0,1)
Teacher is 30 to 39 years old	Dummy (0,1)
Teacher is 40 to 49 years old	Dummy (0,1)
Teacher is 50 to 59 years old	Dummy (0,1)
Teacher has less than 6 years of teaching experience	Dummy (0,1)
Teacher has 6 to 15 years of teaching experience	Dummy (0,1)
Student completes class worksheets more than once a week	Dummy (0,1)
Student answers questions in class more than once a week	Dummy (0,1)
School socio-economic status [b]	Continuous (mean = 0, s.d. = 1)
Serious absenteeism problem at the school	Dummy (0,1)
Moderate absenteeism problem at the school	Dummy (0,1)
School located in an urban area	Dummy (0,1)
School located in a sub-urban area	Dummy (0,1)
School offers extended learning time to more than 75% of learners	Dummy (0,1)
High parent involvement at school [c]	Dummy (0,1)
No free or subsidised lunch programme offered	Dummy (0,1)
Class size larger than 30 learners	Dummy (0,1)
Reading series used in classroom teaching	Dummy (0,1)

Table A3.1 continued: Descriptions of control variables

Variable	Description
Long books with chapters used in classroom teaching	Dummy (0,1)
High level of teacher collaboration	Dummy (0,1)
Teacher gives reading homework weekly	Dummy (0,1)
Teacher uses worksheets in classroom teaching weekly	Dummy (0,1)
Teacher uses group discussion in classroom teaching weekly	Dummy (0,1)
Teacher asks learners to give oral feedback of reading weekly	Dummy (0,1)
Diagnostic testing emphasised in classroom	Dummy (0,1)
Teacher has at least a university degree	Dummy (0,1)
Teacher has a post-matriculation diploma	Dummy (0,1)
Western Cape province	Dummy (0,1)
Northern Cape province	Dummy (0,1)
Free State province	Dummy (0,1)
Kwa-Zulu Natal province	Dummy (0,1)
North West province	Dummy (0,1)
Gauteng province	Dummy (0,1)
Mpumalanga province	Dummy (0,1)
Limpopo province	Dummy (0,1)

[a] PIRLS generated variable

[b] Calculated as the average socio-economic status of students in the school.

[c] Parent involvement is coded as taking a value of 1 if the school has more than two formal parent-teacher conferences per year and parents are actively involved in school; 0 otherwise.

Chapter 4

Balancing Act: A Semi-parametric approach for determining the local treatment effect of school type with an application to South Africa

Despite the abolishment of a racially segregated schooling system in 1994, schools that principally served white students under apartheid remain functional and those that served black African students remain dysfunctional and largely incapable of producing results. The link between socio-economic status (SES) and performance continue to define the South African schooling system. This study adds to the evidence through estimating the causal (treatment) effect of attending an English/Afrikaans testing schools where the language of instruction at the school serves as a proxy for former department. A recently defined class of balancing weights by Li, Morgan and Zaslavsky (2014) are used in conjunction with non-parametric coarsened exact matching to calculate the local treatment effect for the sample of students (and schools) with optimal overlap. Using the full sample, this is estimated to be approximately 18 months of learning. This estimate is not significantly changed when the sample is restricted to those schools with similar distributions of inputs targeted directly by government policy.

4.1 Introduction

The question of whether one type of school produces better educational results than another type of school is central to school effectiveness research. The South African education system is one that reflects deeply entrenched social inequalities driven by a set of highly diverse and unequal institutions that vary greatly in their effectiveness and ability to produce student outcomes. Evidence hints towards a "bimodal" distribution of student performance; that is, a different data generating process for historically advantaged schools than for historically poor, predominantly black, schools (Gustafsson, 2007; Van der Berg, 2007; Fleisch, 2008; Taylor, 2011; Spaull, 2012). Before 1994, the schooling system was one divided into fifteen education ministries: a ministry for the central planning of national norms and standards; four racially defined school departments (for black Africans, coloureds, Indians and whites); and ten Bantustan (homeland) departments. Each

school department had its own school models and funding formula. This ultimately led to significant disparities in the type and quality of education received and school functioning that have persisted despite the conglomeration of schools under a single National Department of Education. As noted by Spaull (2012: 12) "Although the formal schooling institutions of apartheid were abolished… the informal schooling institutions inherent in non-white schools remained largely intact", including but not limited to socio-economic disadvantage, a lack of strong school leadership and efficient management, lack of parent involvement and poor discipline on the part of students and teachers.

Despite a lack of conclusive evidence of the school resources and settings that are predictive of better outcomes (Hanushek, 1986), policy remains focused on equalising performance through expenditure aimed at equalising resources. According to Motala and Pampallis (2005), the school finance literature offers five definitions of resource equity in school inputs that are relevant to the South African context: equal opportunity, wealth neutrality, horizontal equity, vertical equity and adequacy. Unequal educational opportunities in South Africa remain a great obstacle to equality, particularly given the significant role that family characteristics play in determining school outputs. The absence of equal opportunities naturally leads to the concept of wealth neutrality; that is, the quality of education available to a child should not depend on their home circumstances and the wealth of their immediate community (Motala & Pampallis, 2005: 54). Yet in his analysis, Yamauchi (2011) indicates the lasting effects of the spatial segregation policies of apartheid as black African students tend to live further from good schools typically situated in expensive neighbourhoods. Geographic inaccessibility is not the only hurdle faced by poor households as financial inaccessibility results from the higher school fees charged by good schools. Whilst the student bodies of the historically white, Indian and coloured schools tend to be racially diverse, although similar in socio-economic background, former black African and homeland schools remain racially homogenous (Spaull, 2012).

Horizontal and vertical equity forms part of the current approach to addressing inequality in schooling inputs; for example, two goals of the new South African system were to equalise spending per student across provinces and equalise pupil-teacher ratios across schools. However, distributional equity has largely ignored equality of outputs and the role that private funding plays in the equity of inputs, particularly with respect to inputs not under policy control (Ladd & Fiske, 2008). Whilst equity has been achieved in respects of certain school resources through, most notably, the South African Schools Act of 1996 (SASA) and the National Norms and Standards for School Funding (NNSSF), poor schools in poor regions continue to experience resource shortages and poor school governance.

According to the Norms and Standards for School Funding (NNSSF) (2006), the distribution of non-personnel funding within provinces is meant to be pro-poor. Schools are ranked and placed into poverty quintiles based on (i) the poverty of the school community and (ii) school conditions, with the result that resources be allocated based on this school poverty index. The poorest schools (quintiles 1, 2 and 3) are classified as fee-free schools and are meant to receive 80% of the available NPNC funding.[71] From table 1.3 of chapter 1 we can see that whilst most provinces meet prescribed spending levels (or at least the minimum threshold), some provinces are underfunding the poorest schools whilst others are overfunding the wealthiest schools. This not only suggests an inequitable distribution of resources among provinces, but also poor fiscal management by provinces. Insufficient capacity within provincial and district level management to process schools' requests for goods and services have led to late delivery as well as late financial transfers (Taylor, 2010: 22).

It is clear that students within the South African school system are not equal in terms of *inter alia* home wealth, exposure to English, parental education and early childhood development. Schools in which there is a higher concentration of students from, for example, poor and uneducated homes would require more resources in order to provide an *adequate* level of learning. Implementing adequacy in school input provision is complex, not least because adequate education is difficult to define but also because the relative cost of serving large proportions of disadvantaged students would need to be determined (Motala & Pampallis, 2005: 55).

It is now argued that school quality has been reduced to what can be raised through school fees, with good quality education in South Africa linked to the likelihood of residents in the local community being able to afford investments in schooling (Yamauchi, 2011). Schools fees have allowed for the maintenance of higher quality facilities in mainly the wealthier quintile 5 schools with the subsequent movement of children whose parents are able to pay high user fees into these better resourced schools. The result is a yawning gap in resources between rich and poor schools on the one hand, and a yawning gap in performance between rich and poor students on the other.

Private spending in the form of school-fees changes the picture of equalization to one of substantial divergence within the public schooling sector. As was displayed in table 1.2, the non-personnel departmental spending norms that aim for approximately 6 times the expenditure in quintile 1 than in quintile 5 schools is far removed from a reality where, as a result of private funding, spending per student in quintile 5 schools is roughly 3 times that in quintile 1. This is in exact reverse to what the policy intention of creating equity in expenditures hopes to achieve for promoting redress amongst the disadvantaged student population.

[71] This used to be the poorest 40% of schools who were allocated 60% of the available NPNC funding.

The primary aim of the analysis of this paper is to understand the role that inequitable distributions of resources play in determining performance differentials between former advantaged and former disadvantaged schools. The Progress in Reading and Literacy Study (prePIRLS) data collected in 2011 is employed, and the potential outcomes framework from the treatment literature is adopted to define the effect of attending a former advantaged school. The methodological approach involves pre-processing the data through (1) finding a subsample of schools for which sufficient overlap in *pre-treatment* school resources exists and (2) generating balancing weights that are based on the propensity of attending a former advantaged school. Pre-processing the data in this way through breaking (or at least reducing) any linkages between the treatment variable Z and the control variables allows estimates based on subsequent parametric analysis to be far less model dependent (Ho, Imai, King & Stuart, 2007). This paper also illustrates that the use of post-treatment school resources for matching will erroneously bias the estimated treatment effect of school type. This paper further argues that the *local* average treatment effect estimated is for a subgroup of "marginal" students for which the comparison across the two school types is not only most relevant, but also potentially more interesting.

The paper proceeds as follows. Section 2 details the study motivation and methodological approach. Section 3 describes the data employed, followed by a discussion of the empirical results and sensitivity checks in section 4. Section 5 concludes.

4.2 Motivation and Methodological Approach

4.2.1 Study Design: The effect of attending a former advantaged school

In answering questions of school effectiveness with regards to school type attended we might consider an optimal design to be one that randomly assigned students across the two sub-systems. However, such experiments are logistically infeasible. As is the case in many countries, schooling data in South Africa is characteristically observational in nature; that is, treatment assignment is non-random, not controlled by investigator and not known. In order to infer "cause and effect" relationships about attending a former advantaged school, we need to view observational studies as approximations to randomized experiments. This requires a clear description of the hypothetical randomized experiment that led to the observed data.

The standard approach would be to imagine that the treatment assignment (to former advantaged schools) operates on students. However, non-random assignment in observational studies means that there is likely to be a lack of overlap in the covariates, which can lead to significant bias. Balance in covariates is critical in order for a causal comparison between groups to

be made. Keele (2012) makes the conjecture that when interest lies in school effects, the hypothetical experiment should be one that focuses on assignment at the group (school) rather than the individual (student) level. One advantage to group level assignment is that is provides resistance to selection effect. Taking treatment assignment to have occurred at the group level implies that covariate balance first be achieved at the school level before matching on student covariates. In an application to Catholic school treatment, Keele (2012) finds that matching Catholic and public schools on school covariates eradicates the treatment effect; that is, once heterogeneity across school types is removed, there is insufficient evidence that mathematics performance in Catholic schools differs from that of public schools.

The research question posed by Keele (2012) is very similar to that of this paper; that is, what is the impact on performance of attending one school type over another. However, I would argue that, at least in the South African context, more consideration needs to be put into the mechanism of treatment assignment. It is evident that South African students select non-randomly into former disadvantaged and former advantaged schools, leading to quite different distributions of student and home background characteristics across the two school groups. It is also the case that certain school level factors differ because of treatment, and not vice versa. For example, the majority of South African schools today have taken on "no fee" school status, whilst former advantaged schools charge relatively high school fees that have allowed them to continue to offer high quality education.[72] This partly explains why there has not been a flight of middle-class white students out of the public school system into private schools, although there have been dramatic movements of black middle-class students into former HOA, HOR and HOD schools.

It therefore comes as no surprise that the matched school and student sample identified by Keele (2012) is only 11 percent of the original sample. If the presence of certain school resources is an artefact of a higher quality school linked, in part, to better students, parents, teachers and management, then it would be foreseeable that matching on these factors yields an insignificant performance gap; this is unless there was some systematic difference in the efficacy with which these resources were utilised across the two school types. The analysis of this paper recognises that differences in the distribution of certain school resources may result either post-treatment (for example, factors related to good governance such as teacher satisfaction and teacher absenteeism) whilst others are more likely to result pre-treatment. Matching on all school resource variables, both

[72] Fiske and Ladd (2004) oppose the view that fees create a situation in which the quality of education is highly correlated with a community's wealth. However, comparisons of Quintile 4 and Quintile 5 schools makes quite clear the difference that fees can make to affording resources such as computer and science laboratories, school buses and smaller pupil-to-teacher ratios that are likely to contribute to augmented performance. The student body that a school attracts, including the affluence and knowledge-base of parents, can further determine the quality of teacher and school management that a school attracts.

post- and pre-treatment, would not only introduce post-treatment bias into the estimated treatment effect but also dramatically limit the comparative school samples.

This study proposes that in finding suitable comparator groups across the former disadvantaged and advantaged school systems, covariate balance be achieved on pre-treatment school resources and students so that the treatment effect will partly be a function of differences in the distribution of post-treatment school resources and partly a function of differences in school effectiveness across the two systems. This is explored further in section 2.3 of this paper.

4.2.2 Potential Outcomes Framework

Consider a sample of *N* units each belonging to one of two groups defined by the binary indicator variable Z, where $Z = 1$ indicates selection into treatment school and $Z = 0$ indicates selection into control. The sample can therefore be divided into n_T and n_C treatment and control units, respectively. For each unit i we observe an outcome Y and a set of pre-treatment covariates X. Interest lies in estimating the effect of treatment. Letting $Y_i(1)$ be the outcome that unit i would have achieved under treatment and $Y_i(0)$ the outcome that unit i would have achieved under non-treatment, the observed outcome for unit i is given by:

$$Y_i = Z_i Y_i(1) - (1 - Z_i) Y_i(0) \quad [4.1]$$

The treatment effect for unit i is given by $\theta_i = Y_i(1) - Y_i(0)$. Given that the two potential outcomes are not observed simultaneously, $Y_i(0)$ needs to be estimated using either a matching algorithm or a weighting estimator.

The propensity score $e(x)$ is defined as the probability of selection into treatment for a given x, $e(x) = \Pr(Z_i = 1 | X_i = x)$. For purposes of this study, $\boldsymbol{X}$ contains pre-treatment student and home background characteristics. Two important assumptions need to be made in the causal effects framework. First, the conditional independence (ignorability) assumption states that, conditional on $\boldsymbol{X}$, treatment assignment is independent of the potential outcomes:

$$Z_i \perp \{Y_i(0), Y_i(1)\} | \boldsymbol{X} \quad [4.2]$$

The second assumption of overlap ensures that there is common support in the covariate distributions across the two groups and each unit i has a positive probability of receiving treatment:

$$0 < e(x) < 1 \quad [4.3]$$

4.2.3 Inducing randomness: Creating comparable school and student groups

Two commonly used nonparametric strategies for balancing covariates across control and treatment groups are matching and propensity score reweighting. Matching involves linking similar individuals from two groups with respect to confounders according to some distance measure, for example, Mahalanobis distance or nearest neighbour. The causal comparison is then based only on the matched sample. Reweighting, on the other hand, applies weights to the entire sample such that the covariate distribution of the two groups of individuals is "matched", with the comparison then based on weighted outcomes. Therefore, whilst matching is designed to create local balance for a subset of the observed sample, reweighting is designed to create global balance. The literature on weighting (c.f. Rotnitzky & Robins, 1995; Hahn 1998; Hirano et al. 2003; ; Busso et al., 2011; Robins et al. 2012) largely focuses on the Horvitz-Thompson (HT) weights, calculated as the inverse of the probability of an individual being assigned to the observed group (Horvitz & Thompson, 1952).

When faced with the options of matching or reweighting, it is relevant to know which of the two methodologies perform better in finite samples. Using simulations, Frölich (2004) finds the weighting estimator to be the worst of all considered estimators in terms of the mean squared error. Busso et al. (2011: 2) come to quite a dissimilar conclusion and find that "an appropriate reweighting approach nearly always outperforms pair matching" in terms of bias and variance, except in cases where overlap is poor. In dealing with lack of overlap, Crump et al. (2009) have characterised the optimal subsample for estimating the treatment effect using a rough rule-of-thumb that discards those units whose propensity score falls outside the range [0.1, 0.9]. This is equivalent to defining a truncated weight. When normalized as opposed to non-normalized weights are used, reweighting can outperform pair matching on the propensity score. A clear disadvantage of weighting is that it relies more on modelling assumptions made in the analysis stage, especially with respect to the propensity score specification.

Two separate approaches for balancing covariates are adopted for the analysis of this paper, one matching and one weighting. Student covariates are balanced using newly introduced balancing weights by Li et al. (2014); these are known as the overlap weights that are proportional to the propensity of assignment to the opposite group. Application of the overlap weights corresponds to the subpopulation with optimal covariate overlap across control and treatment groups, whilst avoiding extreme counterfactuals. Li et al. (2014) have further shown the overlap weights to lead to a treatment effect estimate that has optimal asymptotic variance among all balancing weights. With regards to balancing at the school level, I adopt the coarsened exact matching (CEM) approach of Iacus et al. (2011). The nonparametric method of CEM is a form of "monotonic imbalance bounding" that has been found to out-perform commonly used matching methods in terms of

decreasing, amongst others, imbalance, model dependence, bias and variance (see Iacus et al., 2011a; Iacus et al., 2011b). These methodologies as well as the estimation approach adopted are discussed below.

Propensity Score Reweighting

As discussed in the introduction, a common objective of observational studies is to evaluate the average difference in $Y_i(1)$ and $Y_i(0)$ where the distributions of $\boldsymbol{X}$ across the two groups are balanced. I assume that the sample density of the covariates, $f(x)$, exists with respect to a base measure μ.[73] Li et al. (2014) define the conditional sample average controlled difference (SACE) for a given x as:

$$\theta(\boldsymbol{X}) = E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x) \quad [4.4]$$

In balancing covariates across the two groups, the target sample needs to be represented by $f(x)h(x)$ where $h(.)$ is some pre-specified function of x. Li et al. (2014) further define a general class of descriptive estimands (the weighted SACE) as the average conditional ACE over the target sample:

$$\theta_h = \frac{\int \theta(dx) f(x) h(x) \mu(dx)}{\int f(x) h(x) \mu(dx)} \quad [4.5]$$

The ignorability assumption implies that the above defined SACE is equivalent to the conditional sample average treatment effect (SATE) in the potential outcomes framework (Rubin, 1974, 1978):

$$\theta(\boldsymbol{X}) = E[Y(1) - Y(0)|X = x] \quad [4.6]$$

Therefore, θ_h is the same as the weighted average treatment effect (Hirano et al., 2003b). Both the SACE and SATE require the overlap assumption as defined by equation (3).

For a given $h(x)$, $\theta(\boldsymbol{X})$ can be estimated through weighting $f_z(x)$ to the target sample using the following weights:[74]

$$\begin{cases} w_1(x) \propto \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)}, & Z = 1 \\ w_0(x) \propto \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{1-e(x)}, & Z = 0 \end{cases} \quad [4.7]$$

Different choices of $h(x)$ lead to different target samples and therefore different estimands and weights (Li et al., 2014). The class of weights defined by equation (7) can be broadly thought of as

[73] Where μ is a counting measure and a Lebesgue measure in the case of categorical and continuous variables, respectively.
[74] The weights are proportional up to a normalizing constant.

the balancing weights because they balance the weighted distributions of the covariates between comparison groups.

Probably the most common choice is $h(x) = 1$, with the target sample being the combined control and treated samples and the weights (w_1, w_0) are the HT weights $\left(\frac{1}{e(x)}, \frac{1}{1-e(x)}\right)$. The estimand of interest is then the SATE for the combined sample. Other common choices of h are $h(x) = e(x)$ and $h(x) = 1 - e(x)$. In the case of the former, the target sample is the treated, the weights are $\left(1, \frac{e(x)}{1-e(x)}\right)$ and the estimand is the average treatment effect of the treated (SATT), or $\theta_{ATT} = E[Y(1) - Y(0)|Z = 1]$. On the other hand, the latter choice of h provides the weights $\left(\frac{1-e(x)}{e(x)}, 1\right)$ and the estimand is the average treatment effect of the controls (SATC) and $\theta_{ATC} = E[Y(1) - Y(0)|Z = 0]$.

In the context of the research question posed in this chapter, the SATC would be the estimand of most interest as it measures the expected effect on reading scores of the movement of a grade 4 student attending a former disadvantaged school into a former advantaged school. Alternatively, we could define the SATE, SATT and/or SATC for a truncated sub-sample as suggested by Crump et al. (2009). In this case, $h(x) = \mathbf{1}(c < e(x) < 1 - x)$ and estimands based on this sub-sample will be local weighted average treatment effects.

The balancing weights and corresponding estimand proposed by Li et al. (2014) are the *overlap weights* and *average treatment effect for the overlapped sample* (referred to as the SATO henceforth), respectively. Setting $h(x) = e(x)(1 - e(x))$ implies:

$$\begin{cases} w_1(x) \propto 1 - e(x), & Z = 1 \\ w_0(x) \propto e(x), & Z = 0 \end{cases} \qquad [4.8]$$

It is immediately evident that this weighting places greater emphasis on units with propensity scores close to 0.5 where overlap between the two groups is the greatest. In practice the SATO may be interpreted as the SATE for the sub-sample (or population) that could have gone to either treatment condition. This interpretation is specifically desirable in policy studies since it is these "marginal" units that have a higher likelihood of being responsive to policy intervention and placing focus on these units is likely to be most informative for estimating programme efficacy and future planning.

A further advantage of the overlap weights is that it leads to exact balance between the treatment and control groups on any covariate included in the propensity score model (see Li et al., 2014). This property is limited, however, to the logit function and a propensity score model that includes only main effects. Estimation of the propensity score through alternative methods, such as probit regression, is also likely to lead to good balance. Also, as bias in the estimated propensity

score may be reduced through the inclusion of higher order terms as well as interactions between covariates, there may be a trade-off between bias and exact balancing. When overlap weighting is based on a fully saturated propensity score model it approaches many-to-many exact matching.[75]

This study uses boosted logistic regression (BLR) modelling to estimate the propensity of attending an EAT school, $e(x)$. A number of papers have proposed machine learning methods for estimating propensity scores over commonly adopted methods (McCaffrey, Ridgeway & Morral, 2004; Schonlau, 2005; Westreich, Lessler & Funk, 2010; Lee, Lessler & Stuart, 2010; Austin, Lee, Steyerberg & Tu, 2012). Results from Monte-Carlo simulations and real data applications have led to the broad consensus that ensemble methods in general perform comparably better than logistic regression in causal inference analysis[76] because it averages over multiple simple "weak" classifiers (Schonlau, 2005). In this way, observations incorrectly classified by the previous classifier are weighted more heavily at each step, with the final prediction being a linear combination of the weighted majority from the full sequence of classifiers (Austin et al., 2012).[77]

Coarsened exact matching

Existing matching methods typically comprise of two steps. First, units that fall outside of the common empirical support of both groups are discarded. Second, treated units are matched to control units that are close by some metric. At this point the covariate imbalance can be checked, the matching algorithm re-specified, the imbalanced rechecked, and so forth. In some cases the second step might precede the first if no matches exist for some of the treated. This re-specify-match-check process can become exacerbated when improving balance on one variable comes at the cost of reducing balance on other covariates. In reality, the usual approach to matching skips the "check and re-specify" steps altogether. CEM differs from other matching methods in that the degree of balance is chosen ex ante and the number of matches ex post.

CEM is applied to school level covariates represented by $\boldsymbol{R}$. Specifically, $\boldsymbol{R}$ contains those inputs that have come under direct focus of policy and legislation in terms of redressing the unequal distribution of resources across schools created during apartheid. These include: teacher qualifications, shortages of buildings and shortages in learning and teaching support materials (LTSMs). National benchmarks for these variables are indicated in table 1. Class size and student-

[75] This underpins the bias-variance trade-off inherent to estimating treatment effects; that is, the more complex the propensity score model the higher the variation in weights, whereas the more limited the propensity score model the higher the bias.
[76] The core results of this study were replicated using logit and probit functions to model the propensity scores with insignificant differences in the estimated treatment effect. However, given the high flexibility of boosted modelling in the specification of the functional relationships between the outcome and the covariates, the analysis continues through applying BLR modelling.
[77] For a more detailed description of boosting see Friedman, Hastie, and Tibshirani (2000).

teacher ratios are also factor which have been emphasised by education policy. However, twenty-seven percent of the South African grade 4 classes surveyed in prePIRLS (2011) did not indicate class size, nor is there information regarding the total number of teachers employed within each school, preventing the calculation of student-teacher ratios. An indicator of classroom overcrowding as determined by the school head is supplied, although this is quite a subjective measure that can vary quite dramatically between schools even with the same number of students in grade 4 classrooms. The sensitivity of the estimated treatment effect to the choice of matching variables is investigated later on in this paper.

Table 4.1: National benchmarks for selected school and classroom resources

Selected Indicator	Description	National benchmark
Basic learning materials	Student has at least one exercise book, a pencil or a pen, and a ruler	100%
Student-teacher ratio	Total number of students in a school divided by number of teachers in the school	40:1
Class size	Average number of students per class	40
Teacher education	Higher education qualification	Minimum requirement: four-year teaching degree OR three-year degree with an Advanced Diploma in Education

Source: DBE (2006, 2009, 2011)

The central idea behind CEM is to avoid the curse of dimensionality by provisionally coarsening each covariate into fundamentally meaningful groups. For example, years of teaching experience might be coarsened into less than 5 years, 5 to 10 years and more than 10 years. Exact matching is then performed on the coarsened data with each unit being placed in a single stratum $s \in \boldsymbol{S}$. Strata with at least one treated and one control unit are retained whilst unmatched units are discarded (or given a weight of zero). Treated units in a given strata are given a weight of 1, whilst control units are assigned a weight equal to the number of treated units in the stratum divided by the number of control units in that stratum, $\frac{m_C}{m_T}\frac{m_T^s}{m_C^s}$, where m_T and m_C are the number of matched treated and control units and m_T^s and m_C^s are the number of treated and control units in stratum s, respectively. Weighted comparisons across the two school types should yield distributions of $\boldsymbol{R}$ that are indistinguishable, with distributional differences in school and teacher covariates not included in $\boldsymbol{R}$ supposedly related to treatment. The ultimate goal of matching estimators is to reduce matching error driven by covariate imbalance between groups on the one hand and model dependence on covariates given treatment on the other, whilst at the same time reducing bias and variance. CEM has been shown to eliminate imbalances including nonlinearities,

interactions quantiles, moments and so forth, which in turn avoids model dependence (Iacus et al, 2011a).[78]

Iacus et al (2011a) propose measuring covariate imbalance using a L_1 distance that characterises multivariate differences between $\Pr(\boldsymbol{R}|Z=1)$ and $\Pr(\boldsymbol{R}|Z=0)$. Continuing with the potential outcomes framework, consider $m_T \leq n_T$ and $m_C \leq n_C$ to be well-matched treated and control units and $n_T - m_T$ and $n_C - m_C$ unmatched units, respectively. Letting $H(R_1), \ldots, H(R_k)$ represent sets of distinct values generated by binning on the respective covariate, we can construct a multivariate histogram generated by the Cartesian product $H(R_1) * H(R_2) * \ldots * H(R_k) = H(\boldsymbol{R})$. $f_{\ell_1 \ldots \ell_k}$ is the relative frequency for observations belonging to the cell with coordinates $\ell_1, \ldots, \ell_k$ of the cross-tabulation of m_T and similarly $g_{\ell_1 \ldots \ell_k}$ for m_C. The multivariate imbalance measure is then given by:

$$\mathcal{L}_1(f,g) = \frac{1}{2}\sum_{\ell_1,\ldots,\ell_k \in H(R)} |f_{\ell_1 \ldots \ell_k} - g_{\ell_1 \ldots \ell_k}| \qquad [4.9]$$

The value of $\mathcal{L}_1$ is easily interpretable: if f and g do not overlap at all, then $\mathcal{L}_1 = 1$; if the two distributions overlap completely, then $\mathcal{L}_1 = 0$. The size of $\mathcal{L}_1$ therefore provides useful relative information (dependent on data and covariates used). For example, if $\mathcal{L}_1 = 0.6$ this suggests that 40 percent of the density of the two histograms overlap.

An alternative measure of covariate imbalance is the absolute standardized differences in means. This is calculated as the absolute difference in means between the control and treatment groups divided by a pooled standard deviation (before matching), where the pooled standard deviation is calculated as:

$$sd_{pooled} = \sqrt{\frac{(n_C-1)sd_C^2 + (n_T-1)sd_T^2}{n_C+n_T+2}} \qquad [4.10]$$

An absolute standardised difference less than 0.2 is considered as satisfactory balance whilst a value of 0.1 is considered ideal (Cochran & Rubin, 1973).

The amount of bias reduction and efficiency gain that is possible from pre-processing the school sample in the manner described above depends on (i) the distribution of covariates in the control and treatment groups, (ii) the size of the initial bias in these covariates, (iii) the original sample sizes of the treatment and control groups, (iv) the number of matches selected and (v) the correlation between the covariate/s and the outcome (Ho et al, 2007). If balance is improved through matching, the standard error on the treatment effect will fall. However, if the sample size is reduced too much then the reverse could arise. Ho et al (2007: 214-215) offer some guidance for

[78] Some important properties of CEM are that it bounds model dependence and the treatment effect error, as well as meets the congruence principle. See Iacus et al (2011) for a more in-depth discussion of these and other properties.

applying matching in practice. First, if the number of control units is much larger than the number treatment units, then losing control units until their number approaches that of the treatment group will reduce bias without greatly increasing the variance. Second, if n is reduced so much that variance increases, matching will still be advantageous as long as squared bias (and therefore mean squared error) does not increase. Overall, the approach followed by this study is *doubly robust* in that either the matching or overlap weighting fails, the causal estimates will still be consistent (Robins & Rotnitzky, 2001).

4.2.4 Post-matching estimation strategy

Following CEM, there are $s \in \boldsymbol{S}$ strata each with the same coarsened values of a chosen subset of school covariates. Some strata ($\boldsymbol{S}_M$) will contain both treated and control units whilst other strata ($\boldsymbol{S}_{UM}$) will contain only treated or only control units. Discarding units falling in strata $s \in \boldsymbol{S}_{UM}$ results in a *local* estimate of the treatment effect of attending a former advantaged school. As different numbers of control and treated units are contained within different strata, the chosen model needs to weight or adjust for the different stratum sizes (Iacus et al, 2011a). The simplest local sample average treatment effect estimator is either a weighted difference in means between the treated and control groups or a weighted linear regression of Y on Z.

Adjusting for the strata weights from CEM only leads to covariate balance in $\boldsymbol{R}$ but not necessarily in $\boldsymbol{X}$ as the matching procedure ignores balance at the student level. Balancing at both the group (school) and individual (student) levels could be achieved through several approaches. For example, a weighted regression of Y on Z where the weight used is the product of the strata and the overlap weights computed from a propensity score model of attending a former advantaged school estimated on the subsample of units falling within $\boldsymbol{S}_M$. To achieve overlap across student characteristics within matched school strata, the regression is weighted using balancing weights computed from a propensity score model that is estimated within each matched stratum.[79] The regression coefficient on Z then forms the estimated local treatment effect for the overlap sample (local SATO) of attending a former advantaged school. In the analysis that follows, comparisons are made between the sample estimates of the ATE, ATC and ATO for the full sample of students, and similarly for the CEM matched school sample. The robustness of the main results is investigated through a deeper investigation of the relationship between school covariate imbalances and the treatment effect of school type attended.

[79] An alternative approach would be to compute a propensity score model that includes matched strata indicators as controls.

4.3 Data description

The prePIRLS 2011 dataset comprises a nationally representative sample of 15 744 grade 4 students sampled from 342 schools. The assessment consisted of a reading test that tested both reading for literacy experience and reading to acquire and use information. Final scores derived from Item Response Theory (IRT) analyses were scaled to an international mean of 500 and a standard deviation of 100. This variable will serve as the outcome of interest for this study. In addition to the reading test, students, their parents, teachers and school principals were asked to respond to a number of contextual background questionnaires aimed at collecting information regarding *inter alia* behaviour and attitudes around reading at home and in school, classroom teaching practices and school organisation. It is the richness of the contextual instruments that make the prePIRLS 2011 data specifically attractive for this study as the propensity score of treatment (attendance of a former advantaged school) can be modelled as a function of a multitude student and home background factors that may control for any unobservable characteristics that drive selection into school type.

The former school department of each school was not identified in the data, therefore a proxy is needed. Prior research has typically sub-divided schools on the basis of average school wealth; that is, the wealthiest quintile or quartile of schools as a substitute for the former advantaged school system and the poorest 75 to 80 percent as a substitute for the former disadvantaged, largely black African, school system (Van der Berg, 2007; Taylor & Yu, 2009; Spaull, 2013). However, part of the analysis of this study balances on school SES and therefore this approach would not be appropriate. Instead, the test language was used to identify the two school sub-systems.[80] As in the PIRLS 2006 study that sampled on Grade 5 students, Grade 4 students in the prePIRLS 2011 were tested in 11 of the official South African languages. The test language was selected based on the language of teaching and learning (LoLT)[81] adopted in the foundation phase of learning at the school.[82]

Stratification by language resulted in 73 percent of the school sample testing in an African language, with the remaining 27 percent testing their students in either English (20 percent) or Afrikaans (7 percent), respectively (van Staden & Bosker, 2014). Selecting all schools that tested in

[80] The National School Effectiveness Study (NSES) that assessed Grade 3, 4 and 5 students in literacy and numeracy is the only nationally representative study for which the former department of the sampled schools is identified. However, the student and home background instruments lack depth which limits the efficacy of the methodological approach of this paper.

[81] This implies that students were tested in the language that they had been exposed to at school, which is not necessarily the same as their home language; approximately two-thirds of students were tested in the language they reported to use most often at home.

[82] Foundation phase (FP) in the South African primary school system is classified as Grades 1 to 3.

either English or Afrikaans - henceforth referred to as EAT schools - as being representative of the former advantaged school system would not be wholly correct as a number of former DET and Homeland schools choose to teach in English or Afrikaans during the FP. For example, 9.4 percent of the former DET and Homeland school collected by the National School Effectiveness Study (NSES) reported their FP LoLT as English or Afrikaans. In order to address the issue of overlap between the two school sub-systems in terms of language of testing, a further restriction was applied to the sample of EAT schools. If more than 65 percent of the grade 4 sample from a particular school was found to not speak the test language on a regular basis, this school was dropped from the group of EAT schools.[83, 84] The final sample adopted by the analysis consists of 1 691 Grade 4 students in 44 EAT schools and 11 160 grade 4 students in 231 African language testing schools, henceforth referred to as AT schools. This corresponds with the statistics provided in the national Education Management Information System (EMIS) database of the Department of Basic Education where approximately 15 percent of primary schools are classified as former HOA, HOD and HOR and 85 percent of schools are classified as former homeland and DET.

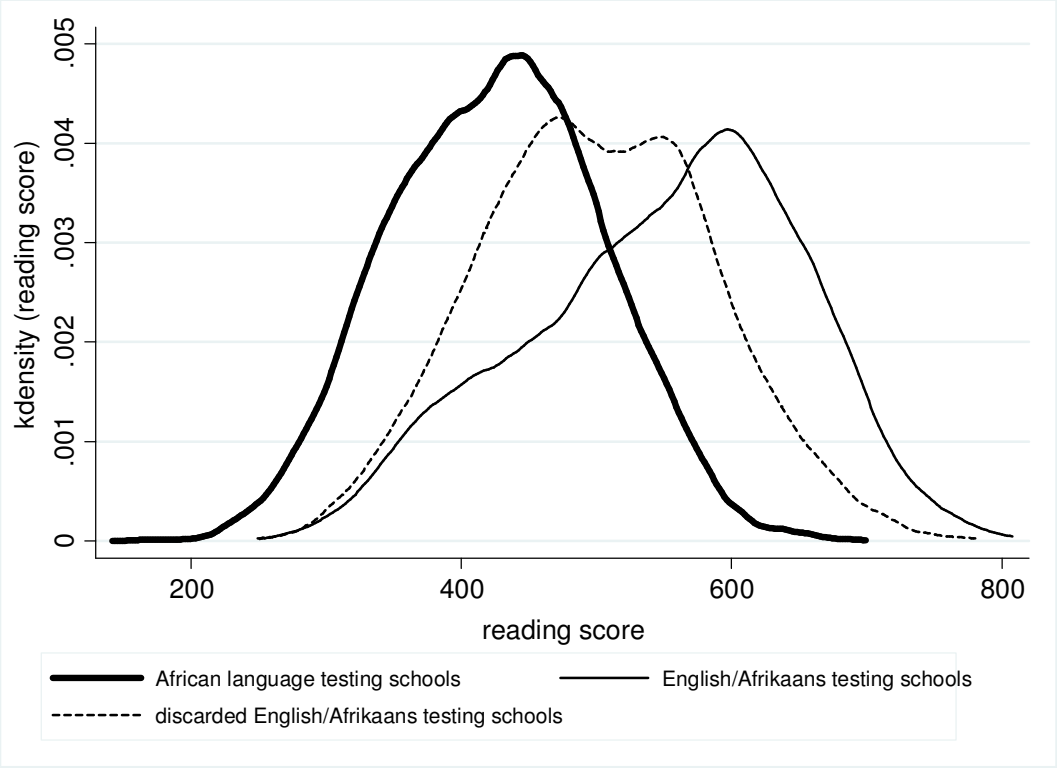
Figure 4.1 depicts dramatic differences in the performance on the Grade 4 reading test across the EAT and AT schools. The score distribution of the group of EAT schools that were discarded for analysis purposes are also indicated. It is clear that this group of EAT schools are a worse performing subset, although there is still substantial overlap with the retained EAT schools. Whilst students attending EAT schools scored a sample average of 532 points in the reading test – close to the high international benchmark - students attending AT schools scored more than an international standard deviation lower (average of 426 points). This gap of 106 points is roughly equal to 2.5 grades of learning (Filmer, Hasan, & Pritchett, 2006).[85]

A number of variables were chosen as potential controls for the propensity score and CEM models. These include 40 student and home background characteristics, 25 school level variables and 25 teacher and classroom level variables. Student and home background controls include

[83] This proportion concurs with the NSES 2008 dataset of Grade 4 students attending former DET and Homeland schools whose FP LoLT was either English or Afrikaans.
[84] Schools were dropped from the analysis as it could not be guaranteed that schools meeting the restriction were, in fact, historically black African schools. It should be kept in mind that the remaining sample of English/ and Afrikaans testing schools may therefore be a sub-sample of better performing former advantaged schools.
[85] In a similar analysis of differences in reading scores across English/Afrikaans testing schools and African language testing schools in the PIRLS 2006 data (Shepherd, 2013), the gap was observed to be 90 points larger than the difference observed here. The PIRLS 2006 survey sampled on Grade 5 students, therefore the larger gap might be explained by a widening of the gap driven by students from poorer schools falling even further behind. An alternative reason may be that, whilst a substantial gap remains, improvements (albeit small) in the poorer school sub-system may have led to a slight narrowing of the gap.

information on household socio-economic status (SES),[86] age and gender of the student, home language, reading activities at home, reading homework activity, books at home (including books in the test language), early childhood development activities and parent employment and education. School level controls capture information regarding average school wealth (school SES)[87], overcrowding in class rooms, student and teacher absenteeism, presence of a school library, frequency of parent-teacher conferences and parent support and involvement, length of school day, textbook shortages and management tasks. Finally, teacher and classroom controls include information regarding curriculum understanding and implementation, teacher collaboration and satisfaction, teacher qualifications, teacher age and experience, classroom teaching practices (including the use of textbooks) and time spent on reading related activities.

Figure 4.1: Reading test score distribution by school test language



Notes: own calculations using prePIRLS (2011)

[86] Household SES is measured by a first principal component analysis of 11 assets that are present in the household including a computer, desk, books, child's own room, internet, newspaper, cellphone, calculator dictionary, electricity and running tap water.
[87] School SES is measured by the average SES of the students sampled within the school.

4.4 Empirical results

4.4.1 Matching students

The BRT propensity score model was fitted in Stata 13 using the *boost* command (Schonlau, 2005). In fitting the BRT model, two parameters need to be specified: the number of splits that will be used for each regression tree, *J*;[88] and the number of iterations, *m*. Hastie, Tibshirani and Friedman (2009) suggest using $4 \leq J \leq 8$. With regards to the number of iterations, too large *m* will result in over-fitting, whilst too few *m* will lead to a poorly fitted model. Regularisation of the BLR model is achieved through a shrinkage factor, λ, and bagging.Shrinking reduces the impact of each additional regression tree in order to avoid model over-fitting, whilst bagging implies that only a random subset of the residuals is selected to build the regression tree at each step. This is thought to reduce the variation of the final prediction without affecting bias (Friedman, 2014). 10-fold cross-validation and a 20 percent test data set were employed in order to assess the predictive accuracy of the propensity score model as well as optimise the choice of parameters *m*, J and λ. In the final BLR model these were selected as *m* = 2 000, J = 4 and λ = 0.005. Forty student and household background characteristics were used as controls in the propensity score model.

Overlaid histogram plots of the estimated propensities of attending an EAT school for both school samples are shown in figure 4.2. The relative propensity score distributions of the two school types provide further evidence of limited overlap in the covariate distributions.[89] An operational drawback of inverse probability weights is that extreme probabilities in the tails lead to potentially explosive weights that can dominate the estimate and lead to a very large variance. Common practice is to truncate the extreme weights based on an arbitrary cut-off point, for example, restricting the region of common support to the propensity score bandwidth [0.1, 0.9] (Crump et al, 2009). The ATO estimand, on the other hand, is able to utilise information from all units whilst avoiding extreme counterfactuals.

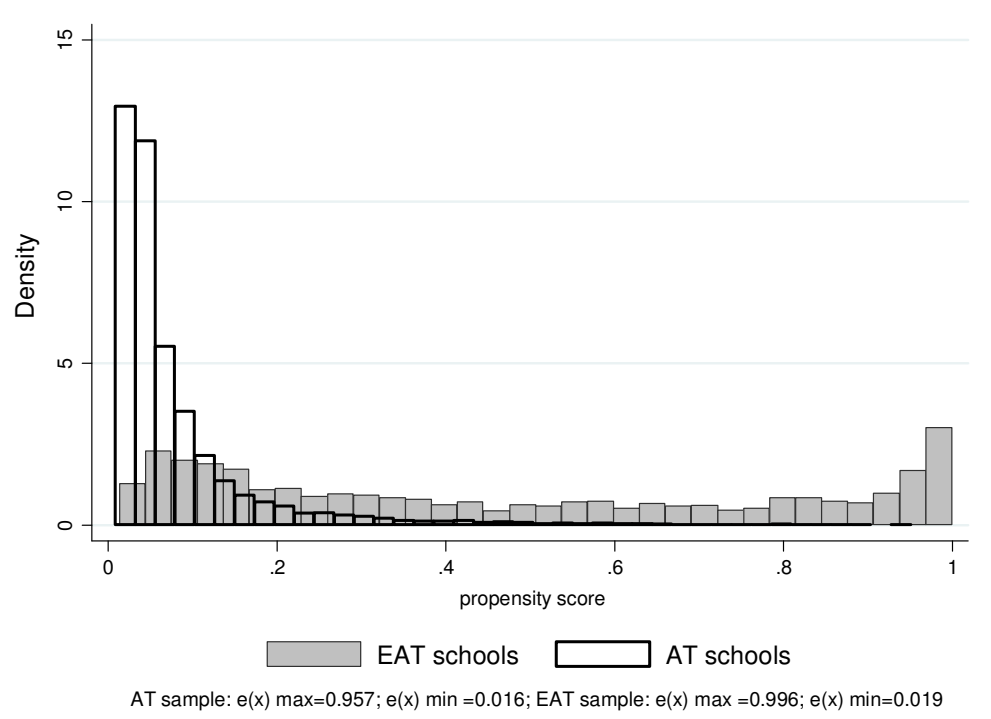
Covariate balance across the EAT and AT groups before and after overlap (ATO) reweighting are illustrated in figure 4.3 and figure 4.4. There is clear evidence of covariate imbalance across the two systems, particularly at the school and teacher/classroom levels where the majority of covariates have a standardised mean difference larger than ± 0.2. This lack of overlap provides a

[88] This defines the number of interactions. Specifying J splits corresponds to a model with up to J -way interactions as J covariates need to be considered jointly. A regression tree with only one split (J = 1) is called a tree stump. Therefore, boosting with stumps fits an additive model, which generally offers a good fit.

[89] Observing the propensity score distributions across control and treatment groups in this way is important, particularly when adopting "trimming" rules for propensity score matching that use maximum and minimum propensities to define the area of common support, as in Dehejia and Wahba (2002). In the context of the current analysis, using this rule would result in none of the units being dropped.

strong incentive for using a reweighting or matching procedure to identify the average treatment of attending an EAT school.

Figure 4.2: Propensity score distribution across school test language



Notes: own calculations using prePIRLS (2011). AT refers to African language testing schools, whilst EAT refers to English/Afrikaans testing schools.

ATO reweighting dramatically improves the covariate balance of student and household characteristics, with at least three-quarters of the covariates meeting the ideal absolute standardised mean difference of 0.1.[90] Despite some improvement, the imbalance in the distribution of school, teacher and classroom characteristics remains after reweighting. Closer inspection of distributional differences across EATS and AT schools reveals similar distributions of school, teacher and classroom factors that are more likely to be under policy guidance. For example, frequency of parent-teacher association meetings, teacher qualifications, formal teaching time (including the proportion of lesson time spent on reading) and the use of textbooks for instruction all have absolute standardised mean differences less than 0.2. However, the distributions of institutional/managerial factors and teacher quality that are more likely linked to within school and classroom processes are vastly divergent; these include the implementation and understanding of

[90] Figure A1 of the appendix illustrates the poorer performance of the other balancing weight schemes in creating covariate balance between the two groups.

curriculum by teachers, teacher collaboration, frequent discussion of parent concerns, teacher absenteeism and the use of advanced teaching aids such as books with chapters.

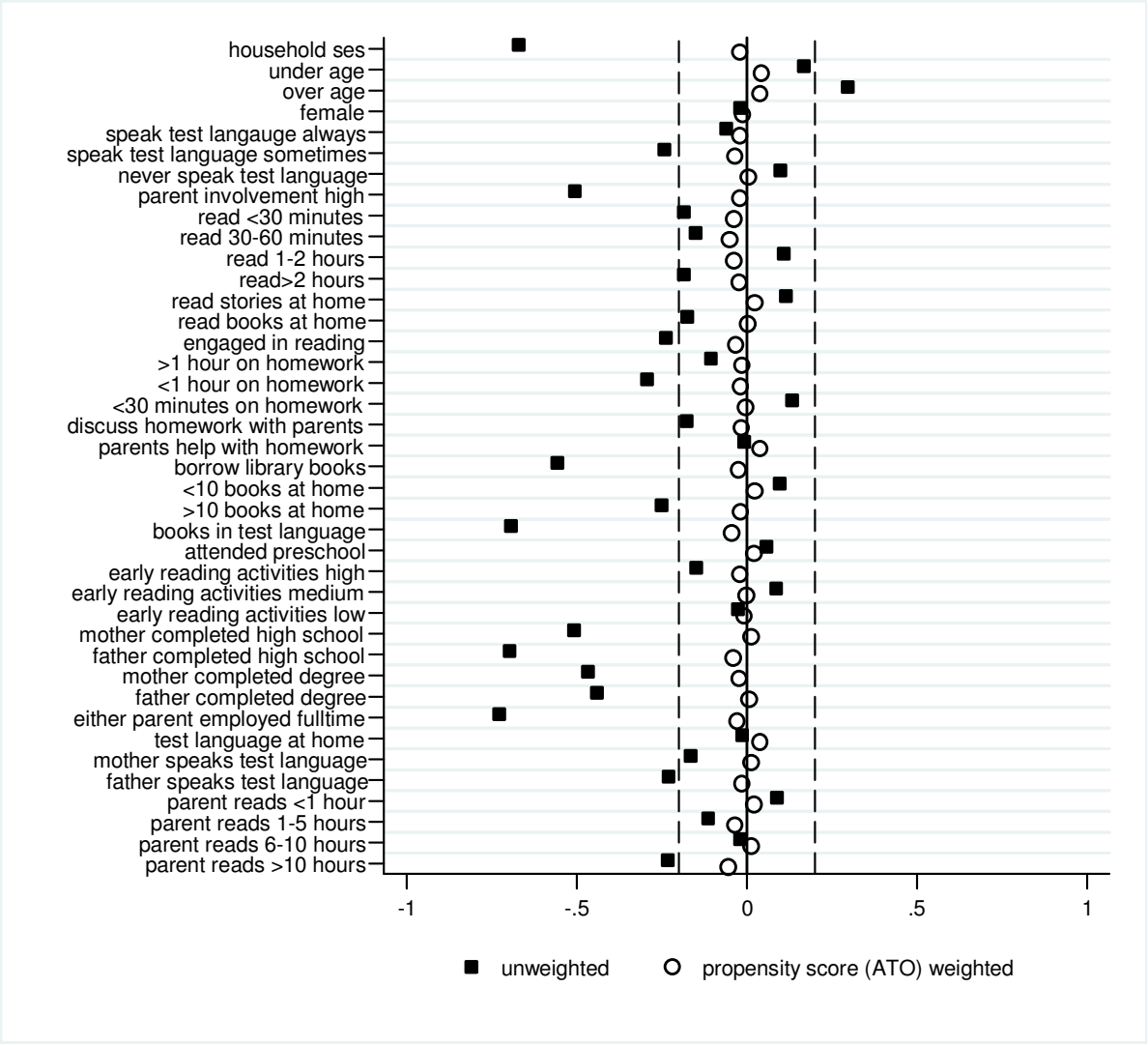
The largest distributional imbalance is observed for school average SES (absolute standardised mean difference of 1.45). From prior research we know school average SES to be a very important determinant of performance in South Africa (Van der Berg, 2007; Taylor & Yu, 2009; Shepherd, 2011; Spaull, 2013). It is not conjecture that wealthier and better educated households are more likely to select into better performing, higher quality schools that are more likely to be attended by children from similarly wealthy and educated households. This is not only because they are likely to be more knowledgeable about the relative quality of different schools, but also because they are able to locate within close proximity of the best schools as well as can afford the higher fees that these schools are likely to charge. A wealthier student peer group not only brings benefits of less social and behavioural problems, but also affords augmented levels of resources such as higher (and better) educated teachers, smaller classrooms and facilities such as computer laboratories and school libraries. We would therefore expect strong positive correlations between the average wealth of a school's students and the presence of high quality school and teaching resources. For this reason, average school SES is often thought to be a proxy for the quality of leadership and overall school culture and learning ethos (McVicar, 2001). The relationship between the treatment effect and average school SES will be explored later on.

4.4.2 Matching schools

CEM is used to match EAT and AT schools with similar covariates so that a comparative subset of schools (and potentially students) can be obtained. As mentioned, I begin by matching on school covariates that represent inputs that have been central to policy in terms of redressing the inequitable distribution of resources across schools, $\boldsymbol{R}$; these include dummy variables indicating shortages of learning and teaching materials (LTSMs) and school buildings, an indicator of classroom overcrowding, indicators of school location and dummy variables for higher education qualifications of teachers.[91] Given that students in South Africa have limited school choice related to location, and that the distribution of teachers and LTSMs are furthermore correlated to the proximity of a school to an urban centre, the location of the school is also used for matching. This model will be referred to as CEM1.

[91] Note that teacher higher qualification is in no way is meant to infer teacher quality, although they may be correlated.

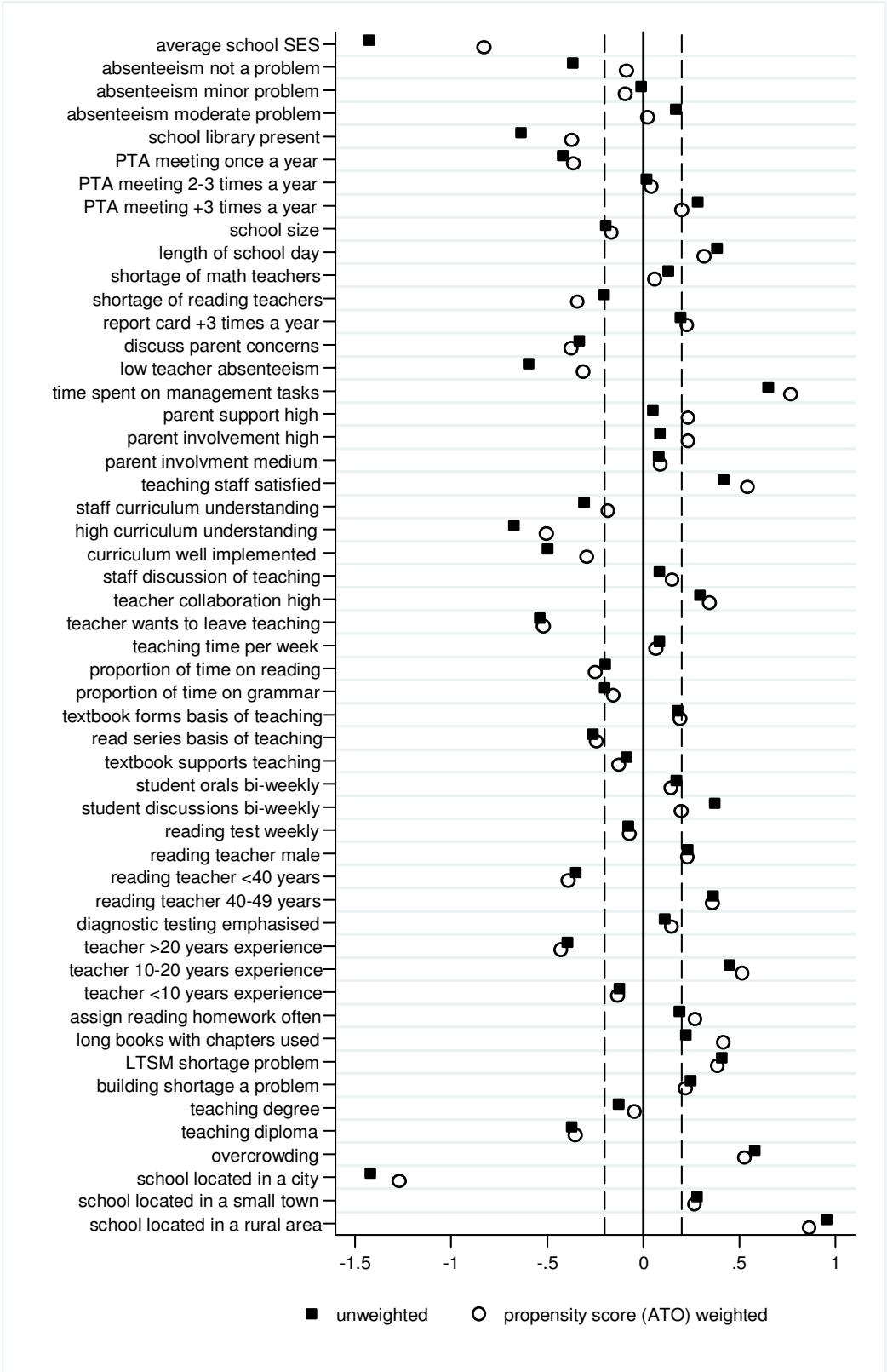
Figure 4.3: Difference in standardised means of student and home background covariates, pre- and post-reweighting



Notes: own calculations using prePIRLS (2011).

Following coarsening and matching, 85 unique strata were created of which 62 contain 129 control (AT) schools only, 8 strata contain 1 EAT school each, and 15 strata contain both treated (36 EAT) schools and control (102 AT) schools. This implies that 38 percent of the original AT sample (4 283 students) is matched to 76 percent of the original EAT sample (1 281 students). Figure 4.5 compares distributions of the remaining school and teacher covariates not included in $\boldsymbol{R}$ that may be related to institutional culture and the quality of management and teaching staff (from this point denoted as $\boldsymbol{Q}$). Comparisons are made across matched EAT and AT schools and matched and unmatched AT schools (comparisons between matched and unmatched EAT schools are ignored given the small sample size).

Figure 4.4: Difference in standardised means of school, classroom and teacher covariates, pre- and post-reweighting



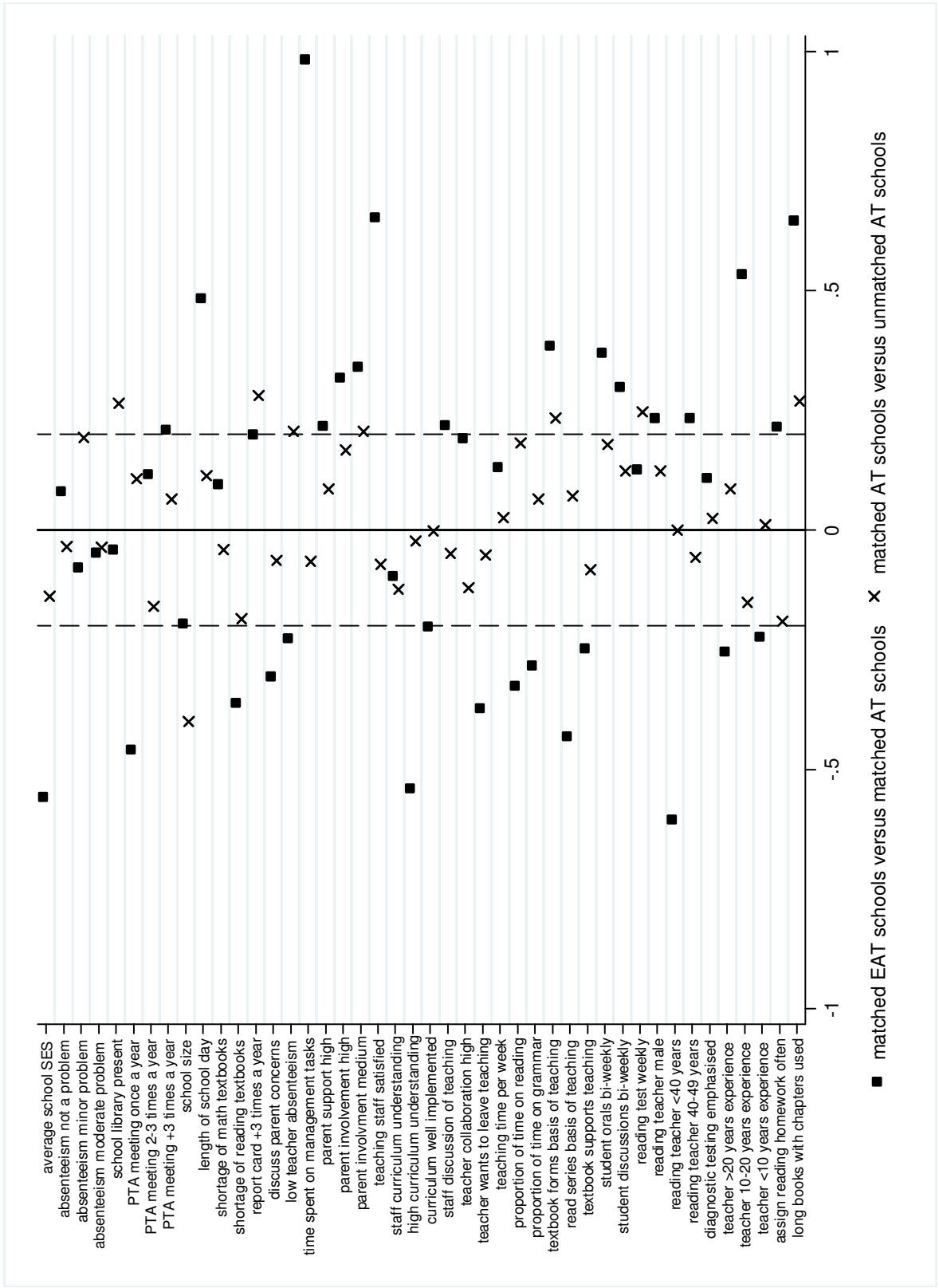
Notes: own calculations using prePIRLS (2011).

It is interesting to note that distributions in $\boldsymbol{Q}$ are less dissimilar when comparing unmatched and matched AT schools; this is not the case when comparing matched EAT and AT school. Therefore, in spite of similar distributions of resources at EAT schools that have been targeted specifically by education policy over the last two decades, the group of matched AT schools are lacking in quality inputs and processes identified by the school effectiveness research (at least within a developing country context) to contribute positively to achievement; for example, the availability of suitably trained and motivated teachers (Smith & Barrett, 2010; Lewin & Stuart, 2003), appropriate textbooks and LTSMs (Barrett et al, 2007; Yu, 2007), improved accountability and parent voice, structured pedagogy that encourages the use of a range of teaching, and learning strategies.

4.4.3 Estimates of the treatment effect

The local SATO estimated from a weighted regression of the treatment dummy on reading test scores is shown in table 4.2. Two estimates of the SATO are compared: a SATO estimated for the matched school sample (column 2); and an estimate that ignores the first-stage school matching procedure (column 1). The SATO estimates are also compared to estimates of the SATC and SATE that adopt the usual inverse probability weights. Although all estimates suggest that there is a significant positive effect of attending a EAT school, the size of the treatment effect differs depending on which weighting method is used. When considering the full sample of grade 4 students, the overlap and truncated HT weights provide the largest treatment effect of attending an EAT school at 63.45 and 62.32 points, respectively. This is compared to a non-truncated SATE of 57.34 points and an SATC of 45.1 points. This suggests that the effect of attending an EAT school is greater for the "marginal" group of students. However, all of the estimated sample treatment effects are not statistically significantly different. Restricting the analysis to schools across the two language settings that are similarly located and have similar distributions of teacher qualifications, LTSMs and building shortages and overcrowding, the ATO, ATC and truncated ATE weights all provide similar estimates of the sample treatment effect of approximately 50 to 55 points. Again, the treatment effect estimates are not statistically significantly different from each other. The larger SATE estimate and standard errors around the SATC and the SATE are likely to be indicative of poor overlap in the propensity scores that results in volatile weights.

Figure 4.5: Difference in standardised means of school, classroom and teacher covariates not included in coarsened exact matching



Notes: own calculations using prePIRLS (2011). AT refers to African language testing schools. EAT refers to English/Afrikaans testing schools.

The results therefore indicate that the different weighting methods provide a relatively identical treatment effect when there is sufficient overlap in covariates, although the SATO is more precisely estimated as it incorporates information from the full sample and places greater weight on the marginal student groups. The fact that the SATO estimate for the matched school sample is smaller, but not significantly so, than that of the full sample indicates that differences in school resources under policy control may play a limited role in driving performance differences across school groups, once accounting for imbalance in student covariates.[92]

Table 4.2: Estimated sample and population treatment effects of attending an EAT school

	(1)	(2)	(3)	(4)
	Sample estimates		**Population estimates**	
	Propensity reweighting only (full sample)	School match and propensity reweighting	Propensity reweighting only (full sample)	School match and propensity reweighting
ATO	63.45***	54.33***	58.01***	54.20***
	(2.99)	(4.33)	(4.00)	(4.76)
ATC	46.94***	53.79***	47.08***	59.87***
	(3.66)	(10.27)	(4.71)	(10.37)
ATE (HT)	53.44***	71.72***	63.91***	79.27***
	(3.61)	(6.62)	(3.84)	(7.43)
ATE (truncated)[93]	59.80***	48.98***	60.91***	51.94***
	(3.91)	(4.50)	(4.82)	(5.12)
Observations	12851	6677	12851	6677
Number of EAT schools	44	43	44	43
Number of AT schools	231	110	231	110

Notes: Treatment effect estimates that incorporate propensity reweighting only employ weights based on propensity score estimates generated from a boosted regression tree model of EAT school attendance. Estimates incorporating school matching and propensity reweighting employ weights based on propensity score estimates (generated from boosted regression tree models of EAT school attendance estimated within each of the 16 strata identified by coarsened exact matching) and matched school strata weights. Bootstrapped standard errors based on 500 repetitions are shown in parentheses. *** significance at the 1% level, ** significance at the 5% level and * significance at the 10% level.

It is useful to give interpretation to the treatment effect in terms of benchmarking it to actual learning. The results of table 4.2 suggest that *localising* the treatment effect to the sample of schools and students with optimal overlap, grade 4 students taught within the former advantaged school system (as proxied by EAT schools) are estimated to perform approximately 0.5 international

[92] From figure 4.3 it can be noted that creating balance on student covariates only already achieves reasonable balance in these five school resources.

[93] Truncation of the sample to include only observations with propensity scores within the bandwidth [0.1, 0.9] results in a sample size of 3231 students without school matching and 2497 students with school matching.

standard deviations higher in the reading test; this is roughly equivalent to 16 months of learning in primary school (Filmer et al., 2006).

4.4.4 Sensitivity analysis

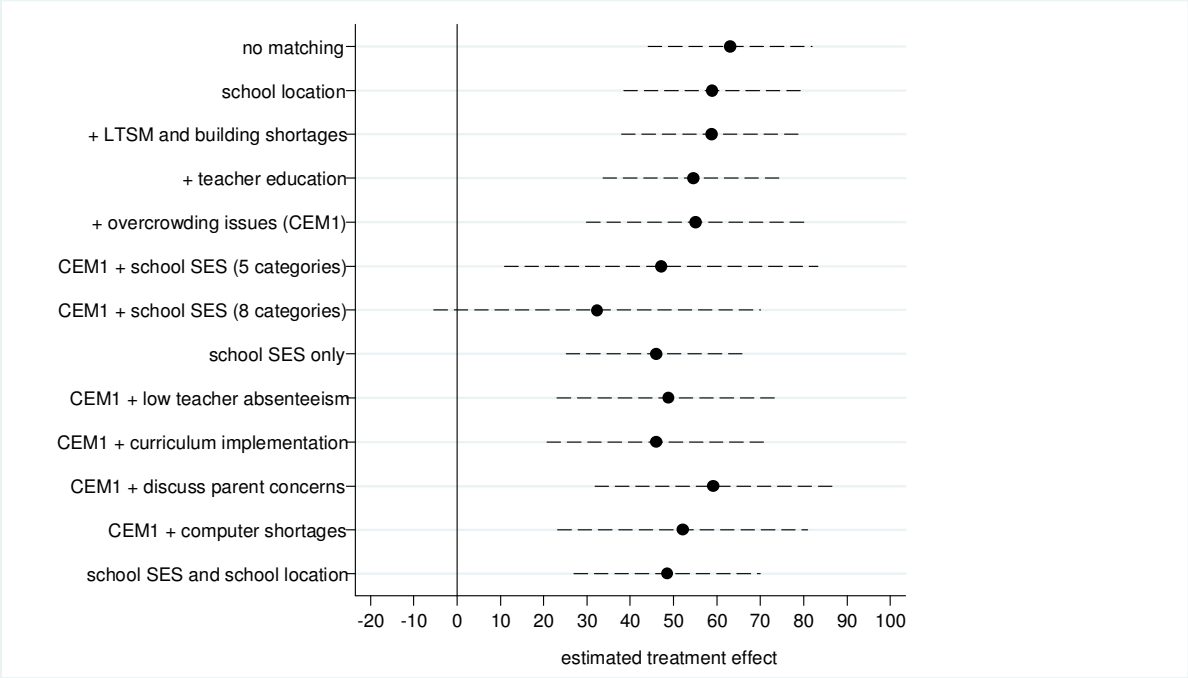
To better understand how correction for imbalances in school resources can change the estimated treatment effect, figure 4.6 presents the estimated SATO as a function of balance on school covariates. Each estimate is calculated after adding an additional school level control to the matching procedure (indicated on the vertical axis). The number of treated and control schools (students) used for estimating the SATO ranges from 44 (1691) and 231 (11160) in the case of no balancing to 18 (728) and 35 (1765) when coarsening is applied to teacher education, shortage problems, class size and school SES using 5 bins. It is clear that stricter balancing reduces the SATO, although this only seems to occur once class size and school SES are added to the matching procedure. In fact, the addition of school SES with the other school level variables (which I will refer to as "CEM1 + school SES") leads to a SATO that is not significantly different from 0. Matching on school SES alone – which may be a catch-all for other school quality resources and school effectiveness –reduces the treatment effect from 63 points (in the case of no matching) to 31 points. Whilst this is not a statistically significant reduction in the treatment effect, a halving of the treatment effect is in no way trivial.

Given that the CEM + school SES leads to a matched sample that is roughly 19 percent of the original sample, a loss in precision is to be expected. In CEM1 and CEM1 + school SES, 16 unique strata containing both matched EAT and AT schools are generated. However, given that the latter results in 53 successfully matched schools, in some instances only 1-to-1 school matches are being achieved, or even 1 AT school matched to multiple EAT schools. The multivariate $\mathcal{L}_1$ statistic for matched schools from CEM1 + school SES is computed to be 0.82; this is not a substantial improvement over the unmatched data $\mathcal{L}_1$ statistic of 0.99 for the same school variables. The matches and strata weights therefore have to correct for quite dramatic differences in the distribution of school SES across the two school groups, likely leading to extrapolation. This is unlike CEM1 which provides a $\mathcal{L}_1$ statistic of 0.23 compared to 0.74 in the unmatched data (accounted for mainly by class size). Extending the CEM1 set of school controls to include alternative indicators of school quality such as low teacher absenteeism and curriculum implementation yields an estimate of the treatment effect that is not dissimilar from CEM1 + school SES (5 bins) that is more precisely estimated. Interestingly, matching on "discussion of parent concerns" increases the treatment effect indicating a negative correlation with EAT language schools, at least within the matched sample; a

higher frequency of discussing parent concerns may be less indicative of greater community involvement and rather points towards a need for more community action.

Given that further school level matching results in imprecisely estimated treatment effects, I investigate the use of regression as (i) an alternative to propensity score weighting and matching and (ii) in combination with propensity score weighting and matching.

Figure 4.6: School treatment effect (SATO) as a function of balance in school level covariates



4.5 Regression meets matching and propensity score weighting

Similar to a doubly robust estimator,[94] regression-adjustment can be used to "mop up" imbalances that may remain between groups (Hill & Reiter, 2006; Ho, Imai, King & Stuart, 2007; Stuart, 2010) as well as increase precision and efficiency and reduce bias (Abadie & Imbens, 2011; Kang & Schafer, 2007; Rubin & Thomas, 2000). It also does not reduce the sample size of interest as in the case of CEM. In a simulation study that compared a regression-adjusted propensity matching estimator to a weighted regression estimator and a doubly robust estimator, Kreif, Grieve, Radice, and Sekhon (2013) showed that regression-adjusted matching is relatively insensitive to misspecifications of both the propensity score (treatment) and the outcome models, even in the presence of unstable weights driven by lack of overlap.

[94] See Bang and Robins (2005).

This study implements the regression-adjusted approach by using the propensity scores to weight Generalised Linear Models (GLMs) written as:

$$F(\mu_i) = \gamma Z_i + \beta' \boldsymbol{X} + \lambda' \boldsymbol{R} + \varepsilon;\ Y_i \sim N(\mu, \sigma^2) \quad [4.11]$$

where $\mu_i = E(Y_i)$ is the expectation of Y_i and F is the identity link function. The regression-adjusted treatment effect estimator is then computed as:

$$\hat{\theta}_{regadj} = \frac{1}{N}\sum_{i=1}^{n}\{\hat{\mu}_i(x_i, Z_i = 1) - \hat{\mu}_i(x_i, Z_i = 0)\} \quad [4.12]$$

where $\hat{\mu}_i(.)$ is the predicted outcome from applying weighted GLMs to the data. Table 4.3 summarises estimates of the SATO across different model specifications that apply (i) regression adjustment, (ii) school matching and (iii) overlap weighting for the full sample (of students and schools) and the CEM1 matched sample. The SATO estimates from table 4.3 are indicated in columns 8 and 9.

Columns 1-5 present the treatment effect estimates for the full sample estimated by regression without any adjustments for school matching or weighting. Controlling for student and home background factors, the treatment effect is estimated to be about 71 points. The addition of the school resources ($\boldsymbol{R}$) included in CEM1 reduces this estimate (albeit not significantly) to 62 points. Controlling further for school SES linear (quadratic) reduces the coefficient on treatment to 50 (40) points. Controlling for all school, classroom and teacher controls (column 5) aside from school SES yields a treatment effect estimate that is no different from column 2.

The results of columns 6-11 present the estimated treatment effect after allowing for propensity reweighting. Columns 6 and 7 are directly comparable with the results of columns 1 and 2, whilst columns 9, 10 and 11 are comparable with columns 3, 4, and 5, respectively. Propensity reweighting and CEM without any further controls yields an estimate of the treatment effect that is approximately 7 points smaller than regression adjustment where weighting is applied simply through the regression estimator. Whilst reweighting might not lead to a statistically significant reduction in the estimated treatment effect, it does suggest that reweighting is able to "mop up" some of the bias induced by poor overlap.

The results therefore indicate that a regression model that attempts to account for selection on observables through fully (or close to fully) parameterising all pre-treatment covariate values (which includes school level variables targeted by policy) provides close to an identical treatment effect as a regression on a reduced sample of schools identified through matching EAT and AT schools. The reason why regression and matching provide similar estimates is because they are in essence both control strategies; the former can merely be understood as a sort of weighted

matching estimator. This is more generally the case when the regression model is (close to) fully saturated-in-X_i. Angrist and Pischke (2008: 51) make the argument that differences between regression and matching are unlikely to be of major empirical importance, as they only differ in the way that weights are used to arrive at the treatment effect. Similarly, there isn't much "theoretical daylight between regression and propensity-score weighting" (Angrist and Pischke, 2008: 63). The primary difference is in implementation, and even without full saturation the use of the right covariates can get you an answer that is close enough to that obtained using propensity scores.

Whilst matching uses the distribution of covariates among the treated to weight covariate-specific estimates into an estimate of the ATT (such as the strata weights constructed through CEM), regression produces a variance-weighted average of the covariate-specific effects. Angrist and Pischke (2008: 56) show that the weights used by the regression estimand are given by:

$$[P(X_i = x|Z_i = 1)(1 - P(X_i = x|Z_i = 0)]P(X_i = x) \qquad [4.13]$$

which imply that regression puts the most weight on covariate cells where $P(Z_i = 1|X_i = x) = 0.5$. It is also worth noting that both the regression and the covariate-matching estimands place zero weight on covariate cells that do not contain both treated and control observations. Angrist and Pischke (2008) further show that the weighting function applied in regression is related to the Horvitz-Thompson ATE estimand. Specifically, the weights are:

$$\begin{cases} w_1(x) \propto \frac{1-e(x)}{E[p(X_i)(1-p(X_i))]}, & Z = 1 \\ w_0(x) \propto \frac{e(x)}{E[p(X_i)(1-p(X_i))]}, & Z = 0 \end{cases} \qquad [4.14]$$

These weights are identical to the overlap balancing weights of Li et al (2014) except that they are normalised by $E[p(X_i)(1 - p(X_i))]$. Therefore, the SATO and the regression coefficient estimated from a regression model that is close to saturated for the covariates should be the same. This is confirmed by the results of columns 1 and 9 which, although not of the same magnitude, are not statistically significantly different. Given that a highly flexible and non-parametric method for estimating the propensity score model was adopted for the analysis, the difference in the estimated treatment effect is due to the fact that the model of column 1 is not as close to saturated as it could be.

Table 4.3: Estimated effect of attending an EAT school accounting for overlap weighting, matching and regression adjustment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Weighting	None					Overlap weighting					
Matching	None					None	CEM1	None			
Treatment effect	70.75***	62.29***	49.63***	40.05***	60.47***	63.45***	55.09***	59.97***	52.91***	37.70***	52.73***
	(8.04)	(8.54)	(7.73)	(7.74)	(8.94)	(10.30)	(12.85)	(6.94)	(8.02)	(7.45)	(8.21)
Regression controls:											
Student/home	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes
School inputs used in CEM1	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
School SES	No	No	Yes	Yes[a]	No	No	No	No	No	Yes[a]	No
School controls (excl. school SES)	No	No	No	No	Yes	No	No	No	No	No	Yes
Average score gap	107	107	107	107	107	107	105	105	105	105	105
R-squared	0.389	0.394	0.416	0.425	0.424	0.118	0.097	0.386	0.394	0.453	0.450
Observations	12851	12851	12851	12851	12851	12851	5560	5560	5560	5560	5560
EAT schools	44	44	44	44	44	44	36	36	36	36	36
AT schools	231	231	231	231	231	231	102	102	102	102	102

[a] School SES is included as a quadratic function

Notes: Unweighted treatment effects are estimated as the partial regression coefficient on a treatment dummy in an OLS regression model of reading scores. The overlap weighted estimate for the full sample employs overlap weights generated from a boosted regression tree model of EAT school attendance. Overlap weighted estimates when school matching is applied employ weights based on propensity score estimates generated from a boosted regression tree model of EAT school attendance that is estimated within each strata. CEM1 matching is applied to teacher education, shortages of LTSMs and buildings and class size. CEM2 is applied to the same covariates as CEM1 with the addition of school SES (wide bins). Bootstrapped standard errors based on 500 repetitions are shown in parentheses. *** significance at the 1% level, ** significance at the 5% level and * significance at the 10% level.

The findings of table 4.3 provide some empirical evidence for "regression as a computationally attractive matching estimator" (Angrist & Pischke, 2008: 51). However, this assumes that, conditional on the individual characteristics of students (and their home backgrounds) and other pre-treatment covariates, school type is independent of learning outcomes. It should be noted that neither regression nor matching are necessarily free from violation of common support. Although in practice they both impose common support, this does not guarantee that covariate cells will contain sufficient numbers of treated and control observations (in the case of matching) or that the regression model will be sufficiently saturated. As a result, both estimators are likely to make use of some extrapolation across cells, leading to bias in the estimated treatment effect. The results of table 4.3 suggest that the treatment effects estimated by regression alone (as in columns 1-5) may suffer some upward bias due to extrapolation driven by a lack of common support. Additional balancing through matching and propensity score weights is able to remove some of this bias.

4.5 Concluding remarks

The quality of schools within the South African public schooling sector is vastly dissimilar and largely defined across racial, socio-economic and regional lines. Education policies under apartheid that favoured minority groups as well as the governance, financing and post-provision policies since 1994 have resulted in the persistently higher performance of former white and affluent schools over the largely dysfunctional former black African and homeland schools. Though much has been done by government in the way of equalising per-student expenditures across provinces as well as creating pro-poor targeting of non-personnel spending, this has yet to reveal itself in terms of improved outcomes amongst particularly former disadvantaged schools. Although home background plays a significant role in determining performance as well as contributing to greater access to better performing schools, it is also important to assess the impact that school type, separate from the effects of home background, has on student performance.

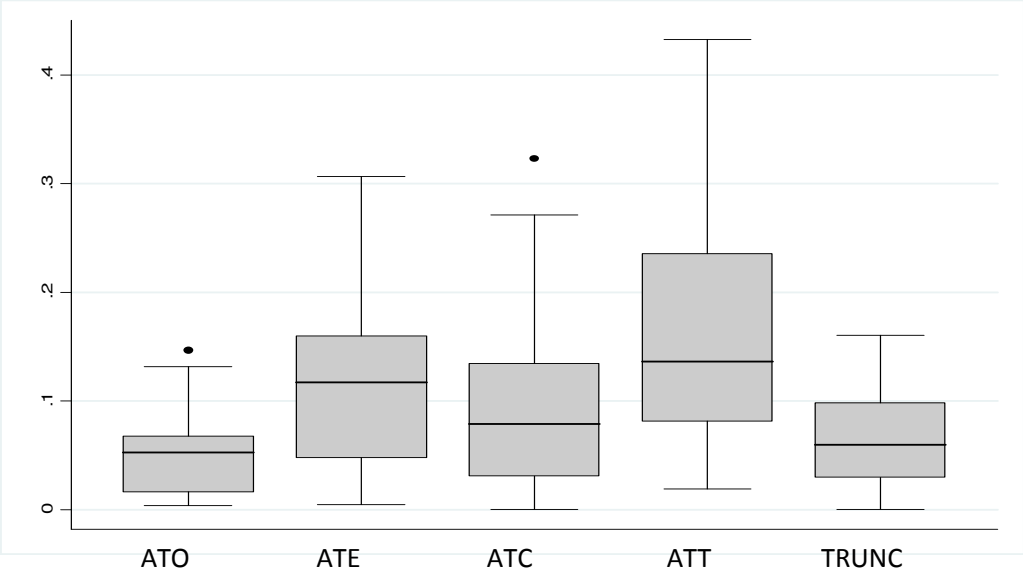
The analysis conducted in this paper made use of the prePIRLS 2011 dataset of grade 4 reading scores, with the schools' language of learning and teaching during the foundation phase of primary schooling serving as a proxy for the former school department. However, given that some (albeit the minority) former disadvantaged schools are likely to teach in English, the treatment effect measured here was that of attending an English or Afrikaans testing (EAT) school. The average reading score gap between EAT and African language testing (AT) schools was 107 points, or roughly 1 international standard deviation that is equivalent to about 1.25 years of learning (Filmer et al, 2006).

Given that the data is observational, the methodological approach adopted had to mimic a random design in order to isolate the (causal) treatment effect of attending an EAT school. This was achieved through the use of a two-stage approach that (1) identified a sample of EAT and AT schools for which sufficient overlap in resources targeted specifically by policy since 1994 exists using coarsened exact matching (CEM) and (2) created overlap balancing weights using estimated propensity scores from a nonparametric boosted regression tree model. The treatment effect was then estimated as the coefficient on a treatment dummy regressed on reading scores using the matched samples of schools and overlap balancing weights. A local treatment effect of approximately half an international standard deviation of attending an EAT school was estimated. This implies that the marginal group of students who attend EAT schools with (1) similar distributions of teacher qualifications, class size, building and LTSM shortage problems and (2) similar distributions of student and home background characteristics as their peers attending AT schools have a learning advantage roughly equal to 12 to 16 months of learning. This is the same treatment effect identified by Coetzee (2014) in her assessment of the treatment effect of an African language speaking student attending a historically white school using a value-added model and instrumental variable regression.

Controlling additionally for school SES as a proxy for school quality factors significantly reduces the treatment effect of attending an EAT school to between 0.2 and 0.3 standard deviations. The results therefore suggest that a significant portion of the difference in performance between former advantaged and former disadvantaged schools is driven by differences in school resources and processes that have not yet been fully addressed by educational policy, and that equity in inputs such as teacher qualifications and class size have had limited effects on closing the gap.

Appendix for chapter 4

Figure A4.1: Absolute standardised differences in student level variables by weighting scheme (CEM1 matched school sample only)



Notes: own calculations using prePIRLS 2011. School matching was performed on five school level covariates using coarsened exact matching. Weights are generated using estimattes from propensity score models estimated within each of the 16 strata containing both treated and control schools. ATO refers to average treatment effect of overlapped sample, ATE refers to average treatment effect, ATC refers to average treatment effect of the controls, ATT refers to average treatment effect of the treated, TRUNC refers to average treatment effect for the sample with propensity scores falling in the range [0.1, 0.9].

Chapter 5

Learn to teach, teach to learn: A within-pupil across-subject approach to estimating the impact of teacher subject knowledge on South African grade 6 performance

This paper assesses the impact of teacher subject knowledge on student performance using a nationally representative dataset of grade 6 students in South Africa. Test scores in two subjects and correlated random error models are used to identify within-pupil across subject variation in performance. Teacher knowledge is estimated to have a positive impact on performance across both the poorer and wealthier subsets of schools once controlling for teacher unobservables. The results suggest that consideration needs to be given to contextual factors such as the quality of teacher training and the working environment within schools and their relationship to the manner in which teacher knowledge is transferred to students.

5.1 Introduction

Almost two decades after the end of apartheid, it is claimed that as many as 90 percent of South African schools "can be labeled as dysfunctional" (Cohen & Seria, 2010). This is in spite of the fact that education gets the biggest share of the country's budget and spending per learner far exceeds that of any other African country. The dismal state of affairs has in part been ascribed to poor teacher education, as well as a broad national concern over the poor state of teachers' knowledge, particularly their subject content knowledge. The President's Education Initiative research project (Taylor & Vinjevoild, 1999) concluded that the limited conceptual knowledge of teachers – including poor grasp of subject - was the most important challenge facing teacher education in South Africa.

Stakeholders in education consider teacher quality to be the most important determinant of student performance. Recent research has shown that variation in teacher quality is a significant determinant of variation in student outcomes (Hanushek, Kain, O'Brien & Rivkin, 2005; Hanushek & Rivkin, 2006). Yet, there is little agreement on what the characteristics of a high quality teacher are, as

well as the relative importance of teacher quality for explaining performance (Hanushek & Rivkin, 2006: 3). Empirical evidence has yet to find strong evidence in support of a relationship between teacher characteristics typically "purchased" by schools - such as a teacher's qualification attained and level of experience – and student achievement. In cases where experience and level of qualification are found to matter, the circumstances tend to be very specific; for example, only the first few years of experience may matter and the effect of teacher qualification may depend on the subject-specificity of the qualification (Goldhaber & Anthony, 2007). Although evidence is somewhat mixed, characteristics such as teacher knowledge and recentness of education are more often than not found to be significantly associated with high student performance in both developed country (Hanushek, 1971; Hanushek, 1986; Monk, 1994; Hanushek, 1997; Wayne & Youngs, 2003; Hill, Rowan & Ball, 2005; Rivkin, Hanushek & Kain, 2005) and developing country contexts (Kingdon, 1996; Mullens, Murnane & Willett, 1996; Tan, Lane & Coustere, 1997; Bedi & Marshall, 2002; Behrman, Ross & Sabot, 2008; Altinok, 2013). The use of, for example, teacher experience and teacher education as policy levers for improving school performance is therefore limited.

The literature has adopted two main approaches to identify the effectiveness of individual teachers in enhancing student performance. These may broadly be classified as value-added or gains models and mixed models. One of the important challenges facing studies attempting to estimate the causal effect of teacher characteristics on student performance is the non-random sorting and selection of students and teachers into classrooms and schools. For example, parents with a preference for achievement will select their children into schools and/or classrooms with high quality, better motivated and knowledgeable teachers. This issue may be addressed through the use of student and teacher fixed effects, although this requires the availability of longitudinal datasets. However, this assumes that students are assigned to teachers on the basis of their time-invariant characteristics rather than time-varying, unobservable characteristics (Ladd, 2008).

This study makes use of a within-pupil between-subject methodology used by Metzler and Woessmann (2012) to estimate the effect of teacher subject content knowledge on grade 6 student test scores in South Africa. This methodology is an extension of the first differencing technique proposed by Dee (2005, 2007) that has been applied quite extensively to eliminate bias from unobserved non-subject-specific student characteristics in order to identify the impact of various teacher and classroom factors such as the teaching style, certification, race and gender of the teacher (Ammermüller & Dolton, 2006; Clotfelter, Ladd & Vigdor, 2006, 2010; Dee, 2005, 2007; Dee and West, 2008; Eren & Henderson,

2011; Lavy, 2010; Schwerdt & Wuppermann, 2011). Identification here relies on variation across teachers in different subjects, as well as student fixed effects across subjects to correct for between and within school sorting of students. This paper adopts a correlated random errors model that allows for the over-identification restriction that is implicit in the fixed-effects model to be tested. We further restrict the sample to students who are taught by the same teacher in the two subjects in order to correct for potential bias due to teacher unobservables.

Two recently compiled case studies in the Gauteng (Carnoy & Chisholm, 2008) and North West provinces (Carnoy & Arends, 2012) of South Africa have provided evidence of a positive relationship between teacher knowledge and student performance. However, stronger positive effects are estimated for quality of teaching,[95] opportunity to learn and teaching institution attended. This study hopes to build on the findings of these studies using the methodology described above and a nationally representative dataset – the 2007 wave of the Southern and Eastern African Consortium for Monitoring Educational Quality (SACMEQ). This dataset is unique in that teachers were asked to complete subject specific tests. To my knowledge, this is the first study to use a nationally representative data set to estimate the effect of teacher subject content knowledge on student performance in South Africa whilst attempting to correct for omitted variable and selection bias. This study also goes further in testing for heterogeneity in the effect of teacher and classroom factors.

The remainder of this chapter is structured as follows. Section 5.2 reviews the literature on teacher knowledge and student performance. Section 5.3 presents the data and basic descriptives and section 5.4 describes the estimation strategy. The main model results and robustness checks are presented in Section 5.5. Section 5.6 concludes.

5.2 Policy context and previous findings in South Africa

The education system inherited by the newly elected democratic government in 1994 was one characterised by high levels of racial segregation and inequality. The general view was that the apartheid curriculum served to prepare black African students with inferior levels of knowledge, understanding and skills in comparison to their white counterparts. The first-ever national audit of teachers in South Africa in 1995 found high numbers of un- and under-qualified teachers as well as fragmented provision of teacher education and training. In attempts to return equality of opportunity to the education system, the current generation of teachers have had to face a number of challenges, including formation

[95] Quality of teaching is measured through classroom surveillance.

of a single national system, the introduction of new curricula and radically changing classroom compositions in terms of language, demography and culture.

The Norms and Standards for Educators (Department of Basic Education, 2000: 47) regarded teachers who had obtained a three-year post-school qualification, or REVQ13,[96] as adequately qualified. The minimum requirement has since been updated to a four-year degree or equivalent qualification (REVQ14) as stated in the 2007 National Policy Framework for Teacher Education. However, a REVQ13 remains to be the norm as an adequate qualification level. In 2004, only 48 percent of teachers met the minimum qualification of a REVQ14. In-service programs offered by universities have allowed teachers to upgrade their qualifications to the necessary level. This is reflected in the rising proportion of annual graduates in Education that are teachers upgrading their existing qualifications. According to the Quarterly Labour Force Surveys (QLFS, Statistics South Africa) of 2010, the proportion of secondary and primary school teachers with REVQ14 and higher was 78.9 and 36.0 percent respectively (68.7 percent together). A further 18 percent are adequately qualified at an REVQ13 level. This implies that in 2010, 13.3 percent, or approximately 55000, of Basic Education teachers remained under-qualified even by the more lenient requirements that applied in 2000.

The quality of content of initial and further training of teachers may vary dramatically given that the current curriculum decisions for pre- and in-service training programs are made independently by individual institutions.[97] Furthermore, the majority of teachers currently in the teaching profession would have received training prior to 1994 when education was racially and ethnically sub-divided and the curriculum was not centralised. A mere 5.4 percent of all practising teachers in 2005 were under the age of 30, which implies that only a limited proportion of teachers are prepared for the new curriculum (Mda & Erasmus, 2008). Some teacher training institutions teach mathematics only up to the level which the teachers would be teaching, which would not provide teachers with an adequate depth of content knowledge or understanding necessary to teach at an Intermediary Phase level.[98] In videotaped observations of mathematics teachers in the Gauteng Province, Carnoy and Chisolm (2008) find that some teachers employ methods that point towards formal training in the use of highly effective methods that require both a deep understanding of the mathematical concepts and pedagogical skills.

---

[96] The Relative Education Qualification Value (REQV) is a relative value attached to an education qualification that is based primarily on the number of recognised prescribed full-time years of study. Completion of school (matric or Grade 12) is an REQV of 10; each additional year of recognized post-school education or training adds one point to the REQV.

[97] At least within the context of the expectations set by the new schools' curriculum and the Norms and Standards for Teachers.

[98] The Intermediary Phase level is defined as grades 4 to 7.

However, the majority of teachers observed were found to use a limited range of teaching methods that were indicative of the rigidity of training received.

Evidence on the impact of teacher knowledge on student outcomes in South Africa is largely unclear. This is mainly due to the fact that teacher subject content knowledge has rarely been captured in large-scale, nationally representative surveys of student achievement. Furthermore, empirical analysis has largely been limited to mathematics. Two recently collated datasets, namely the National School Effectiveness Survey (NSES), a panel dataset covering 3 years of primary schooling, and the 2007 SACMEQ survey provide information on teacher content knowledge through subject-specific teacher test scores. Employing the SACMEQ 2007 dataset to estimate education production functions of student performance, Spaull (2011) finds statistically significant coefficients on teacher content knowledge of 0.074 and 0.048 for reading and mathematics scores, respectively. These estimates are similar to those estimated by Altinok (2013) using multivariate multilevel analysis of the same dataset. These analyses were, however, performed using cross-section least squares methodologies that did not correct for potential bias due to non-random sorting and omitted variables. Additionally, neither teacher education nor teacher experience was included in the regression models; the impact of these teacher quality variables after controlling for teacher knowledge is unclear.

Utilising the NSES panel data, Taylor (2011) finds substantial gains in student learning when teacher knowledge is combined with time on task.[99] However, this only occurs at a very high level of knowledge, indicating a non-linear relationship between teacher knowledge and student performance. The strongest finding by Taylor (2011) is the significant positive relationship between student outcomes and curriculum coverage. Reeves (2005) similarly found that opportunity to learn as measured by curriculum coverage was significantly related to student gain scores in mathematics in a sample of 24 schools in the Western Cape Province.

Two recently conducted South African case studies have paid specific attention to the effect of teacher knowledge on student outcomes. Their methodological approaches further account for non-random sorting across and within schools through the use of value-added modelling. In both studies the authors differentiate between two types of knowledge: content knowledge and pedagogical content knowledge. Shulman (1986) distinguishes between these two forms as knowledge as the former being principally obtained through a teacher's formal pre-service training, and the latter referring to the

[99] The shortness of the teacher tests conducted under the NSES (English teachers were given a comprehension test comprising of 7 questions, and mathematics teachers a 5 mark test) means that this survey provides limited, and potentially noisy, measures of teacher knowledge.

manner in which content knowledge is applied for teaching and is typically obtained through practice or highly skilled training programs. The notion of pedagogical content knowledge has gained wide appeal as it links content knowledge and the practice of teaching and arguably has the greatest ties to effective teaching (Ball et al, 2008). However, Shulman (1987) notes that someone who assumes the role of teacher must first demonstrate knowledge of their subject matter before being able to help students to learn with understanding.

Carnoy and Chisholm (2008) attempt to estimate the contributions of various classroom and teaching factors to learning gains in mathematics of Grade 6 students using a sample of 40 schools in the Gauteng Province. The teacher instrument was designed to include questions that provided measures of content knowledge and pedagogical content knowledge. The findings of Carnoy and (2008) indicated that teachers employed at historically African and coloured schools were observed to score lower in both content knowledge and pedagogical content knowledge than teachers employed within Independent and former white schools where student ability is also relatively higher. Only in the case of the two highest levels of student socio-economic status was performance found to be related to teacher knowledge. Pedagogical content knowledge was strongly positively related to the quality of a teacher's training institution, suggesting that the institution of training may have some direct influence on quality of teaching. Conversely, content knowledge was not found to be significantly related to teaching quality. Value-added modelling of student performance indicated a significant positive effect of teaching quality on test score gains and a positive, but statistically insignificant, coefficient on pedagogical content knowledge. A negative, but statistically insignificant, effect of content knowledge was estimated. This may be driven by the fact that students taught by teachers with higher content knowledge may have experienced lower average gains given higher base test scores. It should be mentioned that value added models were only based on a 25 percent sub-sample of students and t is difficult to say whether the results are upwardly or downwardly biased as the original report gives no details as to how this sub-sample compared to the full sample.

A more recent study by Carnoy and Arends (2012) exploits a natural experiment based on the geographical closeness of South-eastern Botswana and the North West (NW) Province in order to estimate the contributions of classroom and teaching factors to student gains in mathematics. Unlike the Carnoy and Chisholm (2008) study that includes schools from different former departments, the sixty schools selected for this sample are all no-fee (i.e. low wealth) public sector schools in the NW. These are likely to have fallen under the former African school department. Teachers from the NW

sample were found to have less content and pedagogical knowledge than their Botswana counterparts. Teacher knowledge was found to have a strong positive relationship to ratings of teacher quality and opportunity to learn in the NW schools. As in Carnoy and Chisholm (2008) and Reeves (2005), teacher quality and opportunity to learn [100] were estimated to have positive and significant effects on gains in mathematics test scores. However, the effect size of teacher quality was small at 0.05 percent.[101] Teacher mathematics knowledge was not significantly related to achievement gains, possibly due to its positive correlation with teaching quality and opportunity to learn.

In summary, the findings in the South African context seem to suggest that teachers with higher content knowledge, specifically PCK, are more likely to be teaching in wealthier schools that are Independent or fell under the former white and Indian school departments. Therefore, correction for non-random selection is necessary in order to identify the impact of teacher and classroom factors. Teacher knowledge has been found to be positively related to factors associated with effective teaching, such as high teacher quality, opportunity to learn and quality of training, but not to teacher qualification.

5.3 Data and Descriptive Statistics

The data used in this study is the third wave of the SACMEQ survey conducted in 2007. Student knowledge in three subject areas - numeracy, literacy and health - was tested using multiple-choice questionnaires and performance standardized to a regional average of 500 points and a standard deviation of 100 points. Of the 15 countries surveyed, South Africa ranked 10th for reading and 8th in mathematics. [102] In addition to testing, a full array of information regarding home, classroom, and school environments was collected, as well as demographic information on students, parents, teachers and principals. Teachers were also required to complete the health test, as well as subject-specific tests in mathematics and English.[103] This is the first nationally representative education survey in South Africa where teachers' subject knowledge was tested.

---

[100] Here opportunity to learn was defined by content coverage (the number of topics taught during the year) and content emphasis (the number of lessons taught on each topic). These two factors of OTL may have both a direct and an indirect (through quality of teaching) association with student learning gains.

[101] In education, when both dependent and independent variables are measures in standard deviations, the coefficient is referred to as the "effect size".

[102] Other countries surveyed were Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, Swaziland, Tanzania, Uganda, Zambia and Zimbabwe.

[103] Although the SACMEQ II questionnaire did contain a teacher-test, due to South African teacher-union objections, South Africa was one of the few SACMEQ countries that did not complete the teacher-test section of the SACMEQ II survey. This being said, in SACMEQ III teachers were allowed to refuse to write the tests, which some of them did.

Although content knowledge may be related to pedagogical content knowledge, for simplicity's sake this study considers the teacher test score to be a measure of the former. Whilst there was some commonality in questions across the teacher and student tests, teachers were required to answer additional "challenging" questions. To account for differences in difficulty across questions, teacher test scores were transformed using the Rasch scaling[104] so to be directly comparable with student test-scores. For purposes of this study, only scores on literacy and numeracy are considered.[105] Altogether 9083 6 grade students were sampled from 392 schools in South Africa. The large size of the dataset makes SACMEQ III highly advantageous for analysing educational outcomes and their determinants in South Africa. This is especially true given the large intraclass correlation coefficient that is typically observed in school performance data in South Africa (Van der Berg, 2007).[106] After accounting for missing data, the final sample is comprised of 6996 students in 325 schools taught in 686 classrooms by 357 reading teachers and 354 mathematics teachers, where 57 teachers were observed to teach the same students in both subjects.[107]

Table A5.1 of the appendix to this chapter reports descriptives of the final sample. Both the student and teacher scores have been standardised to have a mean of zero and standard deviation equal to one. The estimated model coefficients are therefore expressed as the effect size, or a standard deviation of student performance per standard deviation of teacher subject knowledge. We can compare the estimated effect size to an international benchmark which equates an average learning gain from one year of primary schooling to roughly 30-50 percent of a standard deviation of student achievement (Hill, Bloom, Black & Lipsey, 2008). On average, students performed better in the numeracy test than the literacy test. This may be related to the language of the test as all students were required to write both tests in English.[108] Test scores were found to be positively related to borrowing books outside of school, high household socio-economic status and tertiary education of parents. Both students and teachers performed better in classrooms that were in general better resourced. Test

[104] Rasch (1960)

[105] Performance on the health test was not considered for this study as performance was significantly higher than performance in numeracy and literacy, and there was no significant difference in the health test scores of mathematics and reading teachers.

[106] In calculating the required sample sizes, the first and second waves of the SACMEQ survey erroneously assumed that the intra-class correlation (rho) for the group of countries under investigation would be in the range of 0.3 to 0.4. However, the true rho values in South African fall within the range 0.6 to 0.75, resulting in the samples drawn being too small to obtain the desired significance. The third wave was in this respect a major improvement.

[107] A large proportion of the missing data is due to 15 percent of teachers declining to take the subject-specific tests. Controlling for missing teacher test score as a dummy in the analysis does not significantly alter the results presented in this paper. However, is it probable that the teachers who refused to write the tests are likely to be those with poor subject knowledge. This limits the generalizability of the results around teacher test scores.

[108] Given that the scores on the two tests are standardised across all SACMEQ countries, language may only account for a small part of the difference.

performance of teachers and students were further negatively related to strike activity by teachers and positively related to higher teacher qualifications.

Table A5.2 summarises subject-specific differences in teacher and classroom characteristics. In general, teacher and classroom characteristics were fairly similar across the two subjects. Mathematics teachers were more likely to be younger and possessed post-matriculation qualifications, whereas English teachers were more likely to be female, tertiary educated, and had completed more in-service courses in the past three years. Classrooms in which mathematics teachers taught tended to be better resourced, whilst there was a greater availability of textbooks in English classrooms. Further descriptive analysis (not shown here) revealed that girls performed significantly better in both numeracy and literacy, with a larger difference observed for literacy. Teachers with at least a university degree performed better in literacy but not significantly different in mathematics when compared with teachers with only a post-matric but non-degree qualification. When compared to teachers with complete high school or less, teachers with university degrees performed significantly better in both numeracy and literacy.[109] All variables listed in tables A5.1 and A5.2 were included as explanatory variables in the empirical analysis, as well as a set of provincial dummy controls.

5.4 Estimation strategy: correlated random errors model

I consider an educational production function that places explicitly focuses on teacher subject content knowledge:

$$Y_{1i} = \beta_1 Q_{1j_1} + \gamma' T_{1j_1} + \theta' C_{1j_1} + \delta' X_i + \mu_i + \tau_{1j_1} + \varepsilon_{1i} \quad [5.1]$$

$$Y_{2i} = \beta_2 Q_{2j_2} + \gamma' T_{2j_2} + \theta' C_{2j_2} + \delta' X_i + \mu_i + \tau_{2j_2} + \varepsilon_{2i} \quad [5.2]$$

where Y_{1i} and Y_{2i} are test scores of student i in subject s, $s \in (1,2)$ with $s = 1$ and $s = 2$ representing mathematics and reading, respectively. Students are taught by teachers j who are characterized by their score on the subject-specific test Q_{sj_s}, other non-subject-specific teacher characteristics T_{sj_s} and subject-specific classroom characteristic C_{sj_s}. Teacher characteristics besides subject-specific knowledge will differ across the two equations only if a student is taught by different teachers in the two subjects. X_i represents non-subject-specific student (and school) characteristics. The error term is comprised of a

[109] In cases where the same teacher teaches both subjects, classroom controls were subject-variant whilst teacher controls such as age, experience, qualification, strike activity and hours of preparation were subject-invariant.

student-specific component μ_i, a teacher-specific component τ_{sj_s} and a subject-specific student component ε_{si}.

Least squares estimation of β and γ in [5.1] and [5.2] will lead to biased results due to the presence of confounding unobservable teacher and student effects in the error terms. We are able to correct for non-random selection of students into and within schools through conditioning for unobservable time-invariant characteristics of students (such as ability or motivation) that could be correlated with teacher observables including subject knowledge.[110] Following Metzler and Woessmann (2012), the potential correlation of the unobserved student fixed effect μ_i with the observed inputs can be modeled as:

$$\mu_i = \eta_1 Q_{1j_1} + \eta_2 Q_{2j_2} + \kappa_1{}' T_{1j_1} + \kappa_2{}' T_{2j_2} + \lambda_1{}' C_{1j_1} + \lambda_2{}' C_{2j_2} + \phi' X_i + \varpi_i \quad [5.3]$$

The residual term ϖ_{ij} is assumed to be uncorrelated with the explanatory variables. The parameters η, κ and λ are permitted to vary over subjects, but the parameters on student characteristics, ϕ, are assumed to be the same. Substituting [5.3] into [5.1] and [5.2] yields the following reduced-form equations:

$$Y_{1i} = (\beta_1 + \eta_1) Q_{1j_1} + \eta_2 Q_{2j_2} + (\gamma + \kappa_1)'^{T_{1j_1}} + \kappa_2' T_{2j_2} + (\theta + \lambda_1)' C_{1j_1} + \lambda_2{}' C_{2j_2}$$
$$+ (\delta + \phi)' X_i + \tau_{1j_1} + \varepsilon_{1i}' \quad [5.4]$$

$$Y_{2i} = (\beta_2 + \eta_2) Q_{2j_2} + \eta_1 Q_{1j_1} + (\gamma + \kappa_2)' T_{2j_2} + \kappa_1{}' T_{1j_1} + (\theta + \lambda_2)' C_{2j_2} + \lambda_1{}' C_{1j_1}$$
$$+ (\delta + \phi)' X_i + \tau_{2j_2} + \varepsilon_{2i}' \quad [5.5]$$

where $\varepsilon_{si}' = \varepsilon_{si} + \varpi_i$.

Equations [5.4] and [5.5] comprise an exactly identified model with correlated random effects that are easily estimable using ordinary least squares. Note that teacher subject-content knowledge in each subject enters both equations. The magnitude of the η coefficients capture the extent to which estimated teacher knowledge effects are biased due to omitted student characteristics, while the β

[110] In panel models where multiple observations per student are observed over *time*, educational outcomes can be explicitly modelled as a cumulative process. In order to avoid biased coefficients on characteristics of teacher quality/effectiveness, one or more lagged test scores should be included in the model to account for the prior knowledge/learning that the student brings to the classroom. An analogous approach in the context of a cross-subject model would be to represent a student's knowledge at the beginning of the school year through subject-specific test scores taken prior to the beginning of the period of instruction (Clotfelter et al., 2010). Initial test scores of students are not available in the case of this study. Therefore, we make the assumption that a student's initial knowledge in a subject is negligible and any overall ability will be captured by the student fixed effect.

coefficients represent the structural effect of teacher subject knowledge (Metzler & Woessmann, 2012). Following estimation of the above correlated random errors model, the implied effect of teacher subject knowledge on test performance, β_s, is calculated as the difference between the estimated coefficient on Q_{sj_s} in the equation of student test performance in subject *s* and the estimated coefficient on Q_{sj_s} in the equation of student test performance in the other subject.

This model specification allows us to test the over-identification restrictions implicit in fixed-effects models (Ashenfelter & Zimmerman, 1997). The within-student across-subject estimator by Dee (2005) implicitly assumes that teacher effects are the same across multiple subjects. This makes the model over-identified. Following estimation of equations [5.4] and [5.5] it is straightforward to test whether $\beta_1 = \beta_2 = \beta$and $\eta_1 = \eta_2 = \eta$. If these overidentification restrictions cannot be rejected, we can specify a model that equates the β and η coefficients across equations [5.4] and [5.5] which, given $\lambda_1 = \lambda_2$ and $\kappa_1 = \kappa_2$, will yield the conventional fixed effects model that eliminates bias from student unobservables through differencing within students, across subjects. This illustrates that unrestricted reduced-form estimates for the correlated random effects model will always allow the estimation of the fixed effects model.

The above model specification does not prohibit the possibility of student sorting between subjects. Any unobserved subject-specific student characteristics (such as subject-specific proclivity for performance) will be captured in ε_{si} and any unobserved teacher characteristics that may be related to teacher test score will be captured in τ_{sj_s}. For example, unobserved teacher quality may differ in some consistent way between the subjects taught, or students with an aptitude for mathematics may be assigned to teachers with greater subject knowledge.

A direct test of the hypothesis that the relative student ability in the two subjects is uncorrelated with relative teacher subject knowledge is not available for the SACMEQ data. However, the National School Effectiveness Survey (NSES) collected over three years between 2007 and 2009 can be used to infer the underlying relationship. The mathematics and reading scores of a panel of approximately 8400 students in grade 3, grade 4 and grade 5 are observed. As mentioned in section 2, the NSES conducted subject knowledge testing of Grade 4 and 5 mathematics and reading teachers using short multiple choice tests. Although these tests are likely to be imperfect measures of teacher subject knowledge, they will serve for the purpose at hand.

Following the approach taken by Clotfelter et al (2010), I run a regression of student relative ability (measured as the difference between third grade mathematics and reading test performance) on

a dependent variable of the difference between the subject-specific test score of fifth grade mathematics and reading teachers. The model further controls for school fixed effects. Taking student relative ability in reading and mathematics as a proxy for the subject-specific component of the error term, I find that the null hypothesis that there is no relationship between student relative ability and relative teacher subject knowledge cannot be rejected. Therefore, the NSES data provides no reason to question the assumption that the ε_{si} term in a model with student fixed effects is uncorrelated with the explanatory variable of interest. Although subsequent discussion refers to β as the effect of student knowledge, I do not wish to infer causality. Rather, β is a measure of the relationship between subject-specific teacher knowledge and student performance that is not driven by between- or within-school sorting of students.

In order to correct for bias due to unobservable teacher characteristics, I restrict the sample to students taught by the same teacher in both subjects. In this case, $\mathrm{T}_{1\mathrm{j}_1} = \mathrm{T}_{2\mathrm{j}_2} = \mathrm{T}_{\mathrm{j}}$ and $\tau_{1\mathrm{j}_1} = \tau_{2\mathrm{j}_2} = \tau_{\mathrm{j}}$ and the education production function simplifies to:

$$Y_{1i} = (\beta_1 + \eta_1)Q_{1j} + \eta_2 Q_{2j} + (\gamma + \kappa_1 + \kappa_2)'T_j + (\theta + \lambda_1)'C_{1j} + \lambda_2{}'C_{2j} + (\delta + \phi)'X_i + \tau_j + \varepsilon'_{1i} \quad [5.6]$$

$$Y_{2i} = (\beta_2 + \eta_2)Q_{2j} + \eta_1 Q_{1j} + (\gamma + \kappa_2 + \kappa_1)'T_j + (\theta + \lambda_2)'C_{2j} + \lambda_1{}'C_{1j} + (\delta + \phi)'X_i + \tau_j + \varepsilon'_{2i} \quad [5.7]$$

Restricting $\beta_1 = \beta_2 = \beta$ and $\eta_1 = \eta_2 = \eta$ and taking the first-difference of the two equations gives:

$$Y_{1i} - Y_{2i} = \beta(Q_{1j} - Q_{2j}) + \theta'(C_{1j} - C_{2j}) + \varepsilon'_{1i} - \varepsilon'_{2i} \quad [5.8]$$

This specification is equivalent to including student and teacher fixed effects in a pooled regression. Although this specification makes it impossible to identify the impact of subject-invariant teacher inputs such as gender and race, it does eliminate bias from unobservable teacher characteristics variables when estimating the effect of teacher subject-specific knowledge. Due to the limited sample of students taught by the same teacher in both subjects –only 15 percent of the original sample – estimation using this group will serve as a specification check to the main results based on the full sample.

5.5 Results

5.5.1 Base results

In order to provide some continuity with the earlier literature, table 5.1 presents conventional cross-sectional regression estimates based on equations [5.1] and [5.2]. All regression analysis takes the sampling design of the data into account and standard errors are clustered at the classroom level.[111] Standardized test performance in numeracy and reading are used as the dependent variable in all regressions. Given the purpose of the analysis, only coefficient estimates for the variable of interest (teacher subject knowledge) are reported.[112] The OLS specifications presented in columns 1 – 8 control for varying sets of explanatory variables and the final two columns present the results of a seemingly unrelated regression (SUR) that ignores modelling of correlated random errors. The estimates in columns 1 – 4 indicate a significant positive effect of teacher subject knowledge on student test scores in both subjects that is substantially reduced - from 0.43 to 0.175 and 0.132 percent of a standard deviation in mathematics and reading, respectively - after controlling for a full set of student and home background characteristics. The coefficient on teacher knowledge is more than halved after the addition of school, teacher and classroom controls, yet remains statistically significant. There therefore appears to be evidence of (i) substantial correlation between teacher subject-knowledge and observable and unobservable school characteristics and (ii) self-selection of higher quality students and teachers into higher quality schools. Furthermore, even with a fuller set of controls the estimates on teacher knowledge in columns 9 and 10 of table 5.1 are similar to those estimated by Spaull (2011).

Table 5.2 presents the results of the correlated random errors model of equations [5.4] and [5.5]. I begin by estimating a SUR of test performance that allows for the coefficients on all controls across equations [5.4] and [5.5] to vary. Following this, I was able to test for equivalence of coefficients across equations [5.1] and [5.2].[113] The findings suggest that assuming equivalent effects of *T* and *C* across the production functions for mathematics and reading scores may be restrictive, as there is no a

[111] A sampling method of probability proportional to size (PPS) was used to select schools within provinces, and simple random sampling was used to select students within schools. A minimum cluster size of 25 students was randomly sampled from all grade 6 classes in cases where the total number of enrolled grade 6 students exceeded 25; otherwise all students were included in the sample. Clustering at the classroom level accounts for any correlation of errors associated with the common experience of students in a given classroom environment. The inclusion of student fixed effects makes the case for clustering errors at the student level less compelling.

[112] It can, however, be noted that the estimated coefficients on student/family background and school covariates indicate that females perform significantly better on average, as well as students who speak English on a regular basis at home. Mother's education (particularly higher education), household SES, urban school location, community subsidization of teacher, the proportion of non-permanent teaching staff and school SES are significantly positively related to performance.

[113] Results of these equivalence tests are available from the author by request.

priori reason to suppose that the relationship between, for example, teacher qualification and test performance will be the same for both mathematics and reading.[114] The final model specification was chosen such that $\delta$and ϕ are constrained to be the same across the two subject equations, but γ, θ, κ and λ are permitted to vary. The effect of teacher subject knowledge on student performance in mathematics, β_1, is given by the difference between the regression coefficient on the teacher math test score in the math equation and the regression coefficient on the teacher math test score in the reading equation; and similarly for β_2. The results from column 2 indicate a larger positive estimate on teacher subject knowledge in reading than in mathematics. However, the implied coefficients on teacher knowledge in both subjects are not significantly different from zero. Tests of the over-identification restrictions do not reject the hypothesis that the effect of teacher knowledge is the same in both subjects.

Therefore, column 3 presents the results from SUR estimation restricts $\beta_1 = \beta_2$ and $\eta_1 = \eta_2$.[115] The estimate of η in the final restricted model is found to be highly significantly different from zero, indicating positive selection effects. A model specification that failed to account for this would yield an upward biased estimate of the effect of teacher subject-knowledge on student performance. The implied coefficient on teacher subject knowledge predicts that an increase in teacher test scores by 1 standard deviation increase is expected to increase student performance by 1.3 percent of a standard deviation. This result is not significantly different from zero.

5.5.2 Heterogeneous effects across student sub-groups

The majority of students in the South African schooling system are not first-language English speakers. In addition, these students are likely to be taught by teachers who are themselves not first-language English speakers and are from the same ethnic group as their students. This is particularly true for historically African schools. In addition, access to quality schools is often determined by the affluence of a student's home background. The estimated effect from column 3 of table 5.2 may mask heterogeneity in the effect of teacher subject knowledge across different sub-samples of students.

[114] A SUR model that constrains γ and θ to be equivalent across equations [5.1] and [5.2] does not yield significantly different results with regards to the estimated coefficients on teacher same subject (β_s) and teacher other subject test scores (η_s). However, given that this study is also interested in the effect of other observable teacher and classroom characteristics, such a teacher qualification, a model that constrains γ and θ to be the same could lead to erroneous conclusions regarding the returns to these characteristics.

[115] This model is equivalent to estimating a first-difference model that allows for differing coefficients across other teacher and classroom characteristics besides teacher subject knowledge in the two subjects.

Table 5.1: Cross-sectional regressions

	Ordinary Least Squares								Seemingly unrelated regression	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Maths	Reading	Maths	Reading	Maths	Reading	Maths	Reading	Maths	Reading
Teacher test score	0.433***	0.426***	0.175***	0.132***	0.102***	0.059***	0.076***	0.065***	0.064***	0.051***
	-0.028	-0.031	-0.02	-0.022	-0.02	-0.02	-0.02	-0	-0.02	-0.02
Student/home background controls	-	-	X	X	X	X	X	X	X	X
Classroom controls	-	-	-	-	-	-	X	X	X	X
Teacher controls	-	-	-	-	-	-	X	X	X	X
School controls	-	-	-	-	X	X	X	X	X	X
Adjusted R-squared (OLS)	0.18	0.175	0.399	0.508	0.442	0.563	0.461	0.583		
Observations (students)	6996	6996	6996	6996	6996	6996	6996	6996	6996	
Classrooms (clusters)	686	686	686	686	686	686	686	686	686	
Number of schools	325	325	325	325	325	325	325	325	325	

Notes: Dependent variable is the standardized student test score in numeracy and literacy. Robust standard errors adjusted for clustering at class level shown in parentheses. Clustered standard errors in the SUR models are estimated by maximum likelihood. Significance at *** 1% level ** 5% level * 10% level

Table 5.2: Correlated random effects models

	(1) **Unrestricted model**: All coefficients differ over equations (4) and (5)		(2) **Restricted model**: $\delta_1 = \delta_2$, $\phi_1 = \phi_2$		(3) **Restricted model** $\delta_1 = \delta_2$, $\phi_1 = \phi_2$ $\beta_1 = \beta_2$, $\eta_1 = \eta_2$
	Maths	Reading	Maths	Reading	
Implied β_s	0.015	-0.008	0.001	0.021	0.013
χ^2 $(\beta_s = 0)$	0.99	0.20	0.01	1.90	0.99
Prob > χ^2	0.321	0.654	0.940	0.168	0.320
Regression estimates:					
Teacher test score in same subject	0.044**	0.039	0.036*	0.047**	0.044***
	(0.021)	(0.022)	(0.021)	(0.022)	(0.014)
Teacher test score in other subject	0.047**	0.029	0.026	0.035*	0.031***
	(0.019)	(0.020)	(0.018)	(0.020)	(0.012)
$\chi^2(\eta_1 = \eta_2)$	0.32		0.08		-
Prob > χ^2	0.570		0.779		-
$\chi^2(\beta_1 = \beta_2)$	1.32		0.96		-
Prob > χ^2	0.251		0.326		-
Observations (students)	6996		6996		6996
Classrooms (clusters)	686		686		686
Number of schools	325		325		325

Notes: Dependent variable is the standardized student test score in numeracy and literacy. Regressions are estimated using seemingly unrelated regressions (SUR). Implied β_s is calculated as the difference in the coefficient on teacher test score in subject s between the equation of the student test score in the respective subject and the equation of the student test score in the other subject. Standard errors adjusted for clustering at class level shown in parentheses. Clustered standard errors, shown in parentheses and clustered at the classroom level, are estimated by maximum likelihood. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2 of the appendix to this chapter. Significance at *** 1% level ** 5% level * 10% level.

Table 5.3 presents results from estimation of the correlated random effects model for various student sub-groups: students who speak English frequently at home (column 2), students who speak English rarely at home (column 3), students who come from above average SES home backgrounds (column 4) and students who come from below average SES home backgrounds (column 5). Table 5.3 further includes estimates from a model specification that allows for non-linear returns to teacher subject-knowledge through a spline set at above average teacher test scores (column 1). A larger positive effect size of mathematics teacher subject knowledge on mathematics test scores of 0.055 and 0.039 is estimated for the sub-groups of students who speak English often at home and come from high SES backgrounds, respectively. The results of column 1 further provide evidence of a significant non-

linear effect of teacher subject-knowledge on student performance. Specifically, students taught by mathematics teachers who performed 1 (2) standard deviations above average in the teacher math test are estimated to score 6.1 (12.1) percent of a standard deviation higher than students taught by average performing teachers. Similarly, students taught by reading teachers who performed 1 (2) standard deviations above average in the teacher reading test are estimated to score 6.5 (13) percent standard deviations higher than students taught by reading teachers who scored at the mean. Given that English speaking and above average SES students have a higher likelihood of attending former white and Indian schools that (i) perform notably better on average than former African and Coloured schools (see Van der Berg, 2008) and (ii) are able to afford better quality teachers, [116] the results of table 5.3 are believed to provide evidence of potentially divergent effects of teacher subject knowledge across different sectors of the South African primary school system.

The bimodal nature of performance within the South African schooling system is a well-documented finding in the South African education literature (Gustafsson, 2005; Fleisch, 2008; Taylor, 2011; Spaull, 2013). By this it is meant that the overall test score distribution disguises two separate distributions that correspond to two quite divergently performing subsets of the South African school system that are embedded in the formerly separate administration of education for each race group (Fleisch, 2008). Figures 5.1 and 5.2 show the distributions of student and teacher test scores across school wealth quintiles based on school average SES, where the top 20 percent SES schools (Q5 schools) have been separated from the bottom 80 percent (Q1to4 schools).[117]It is clear that the students in the Q5 schools perform more than an international standard deviation (100 points) above the SACMEQ average of 500, whilst students in the poorest schools perform below average. The picture is similar for teacher test scores in that teachers employed within the wealthier subset of schools perform significantly better on average in both subjects. These findings are in agreement with those of Carnoy and Chisholm (2008).

[116] Even though the salary of the teachers a school appoints (the value of which is based on their experience and qualifications) is paid by the state, schools that manage to attract better quality teachers receive larger state subsidies for teacher costs, ceteris paribus. Schools can use fees to appoint additional teachers that may furthermore be of a higher quality.

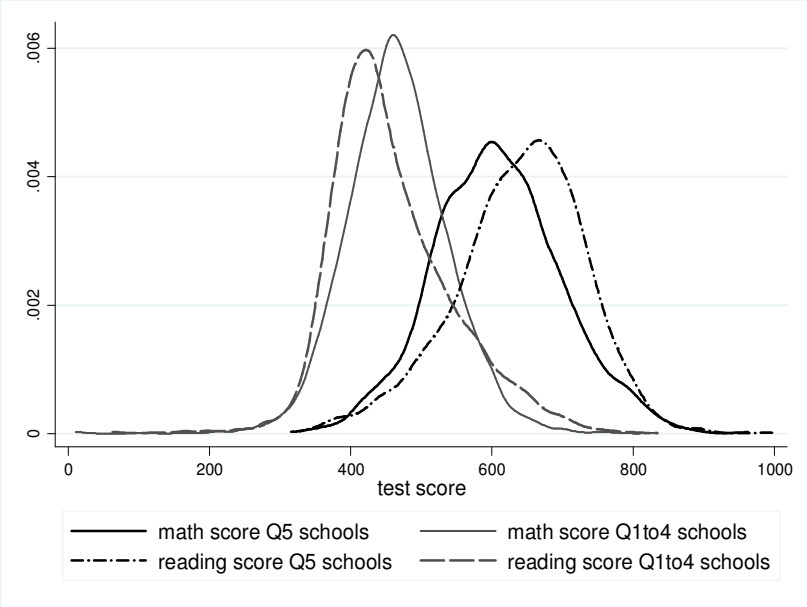
[117] This grouping is chosen based on other studies which have shown no significant difference in performance across the three bottom school SES quantiles (see for example Taylor, 2011; Spaull, 2013). This division is further closely related to the historical separation of formerly black African/homeland schools and formerly white, coloured and Indian schools.

Table 5.3: Correlated random effects models across sub-samples

	Teacher test score level		Student speaks English often		Student speaks English rarely		Household SES above average		Household SES below average	
	(1)		(2)		(3)		(4)		(5)	
	Maths	Reading	Maths	Reading	Maths	Reading	Maths	Reading	Maths	Reading
Implied β_s			0.055	0.017	0.001	0.013	0.039*	0.030	-0.024	-0.008
Prob > χ^2			0.124	0.548	0.949	0.401	0.056	0.104	0.293	0.738
Implied β_s (below average teacher score)	-0.074**	-0.027								
Prob > χ^2	0.010	0.372								
Implied β_s (above average teacher score)	0.061**	0.065**								
Prob > χ^2	0.006	0.021								
Observations (students)	6996		895		6101		3313		3683	
Classrooms (clusters)	686		351		676		646		584	
Number of schools	325		224		323		311		304	

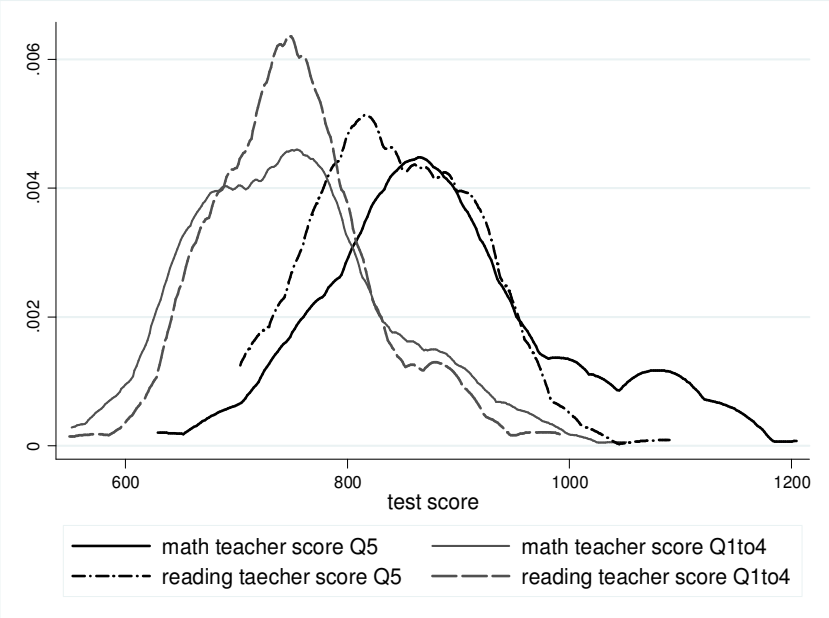
Notes: Dependent variable is the standardized student test score in numeracy and literacy. Regressions are estimated using seemingly unrelated regressions (SUR). Implied β_s is calculated as the difference in the coefficient on teacher test score in subject s between the equation of the student test score in the respective subject and the equation of the student test score in the other subject. In all models, the coefficients on student and school characteristics are constrained, with $\delta_1 = \delta_2$ and $\phi_1 = \phi_2$. Clustered standard errors, shown in parentheses and clustered at the classroom level, are estimated by maximum likelihood. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2 of the appendix to this chapter. Significance at *** 1% level ** 5% level * 10% level.

Figure 5.1: Student performance by school SES quintile



Notes: based on own calculations from SACMEQ III (2007)

Figure 5.2: Teacher performance by school SES quintile



Notes: based on own calculations from SACMEQ III (2007)

Table 5.4 presents the estimated results from correlated random effects models estimated separately for the two school wealth groups. Students test scores across the Q5 and Q1to4 samples have been normalized based on the mean and standard deviation of the respective sub-group. In the case of the Q5 schools, we are able to reject the restriction $\eta_1 = \eta_2$ but not $\beta_1 = \beta_2$ (see column 1). Neither of the over-identification restrictions can be rejected for the sample of relatively poorer schools (see column 3). Using restricted models for each school sample (columns 2 and 4), a significant positive effect of mathematics teacher subject knowledge on student achievement of 11.5 percent of a standard deviation, and a *negative* effect (-0.05) of reading teacher knowledge on student achievement that is not significantly different from zero are estimated. The finding that mathematics and not reading teacher knowledge has an effect on student performance is not surprising given that unlike mathematics, a substantial amount of learning in reading occurs at home.[118] In the case of Q1to4 schools, we find a small negative effect (-0.019) of teachers' subject knowledge that is not significantly different from zero. The estimates for η across the two school samples indicate significant positive selection in Q1to4 schools driven by student unobservables.

The presence of potential non-linearities in the returns to teacher subject knowledge is assessed using a model specification that controls for dummy variables representing teacher test score quintiles defined relative to the school wealth group. Figures 5.3 and 5.4 illustrate the estimated coefficients on the teacher knowledge quintiles across subjects and school wealth samples. The coefficients are plotted against the average test score of the respective quintile and normalized relative to a zero coefficient for quintile 1 of teacher performance. It is immediately clear that irrespective of the ranking of teacher performance, there is no pattern of increasing returns to teacher subject knowledge in Q1to4 schools. Statistical testing confirms that we cannot reject the hypothesis that the returns to teacher knowledge are not significantly different from zero at all quintiles of teacher subject knowledge (see table A5.3 of the appendix). Hence, it cannot be concluded that a student's performance in the poorer subset of schools is significantly better or worse depending on the relative ability of the mathematics and reading teachers. Conversely, the estimates indicate a strong non-linear return to teacher knowledge in Q5 schools. Students taught by the most knowledgeable mathematics teachers perform significantly higher on average, scoring 70 percent of a standard deviation more than students taught by teachers

[118] This is, however, dependent on whether or not learning takes place at home. For example, Spaull (2013) finds that the frequency of speaking English at home and mother's education are positively and significantly associated with reading scores. Gustafsson, van der Berg, Shepherd and Burger (2010) find that the literacy of parents displays a large association with student literacy in South Africa, with the magnitude of parent factors - relative to that of other factors – being arguably larger than is commonly believed.

performing at quintile 1. The returns to reading teacher subject knowledge in Q5 schools rises dramatically when moving from a teacher ranked in the bottom 40 percent of performance to a teacher at the 3rd quintile. Although the returns appear to decline at the 4th and 5th quintiles of reading teacher knowledge, the coefficients are not statistically significantly different from that observed at the 3rd quintile (see table A5.3 of the appendix).

Table 5.4: Correlated random effects models across different school sub-systems

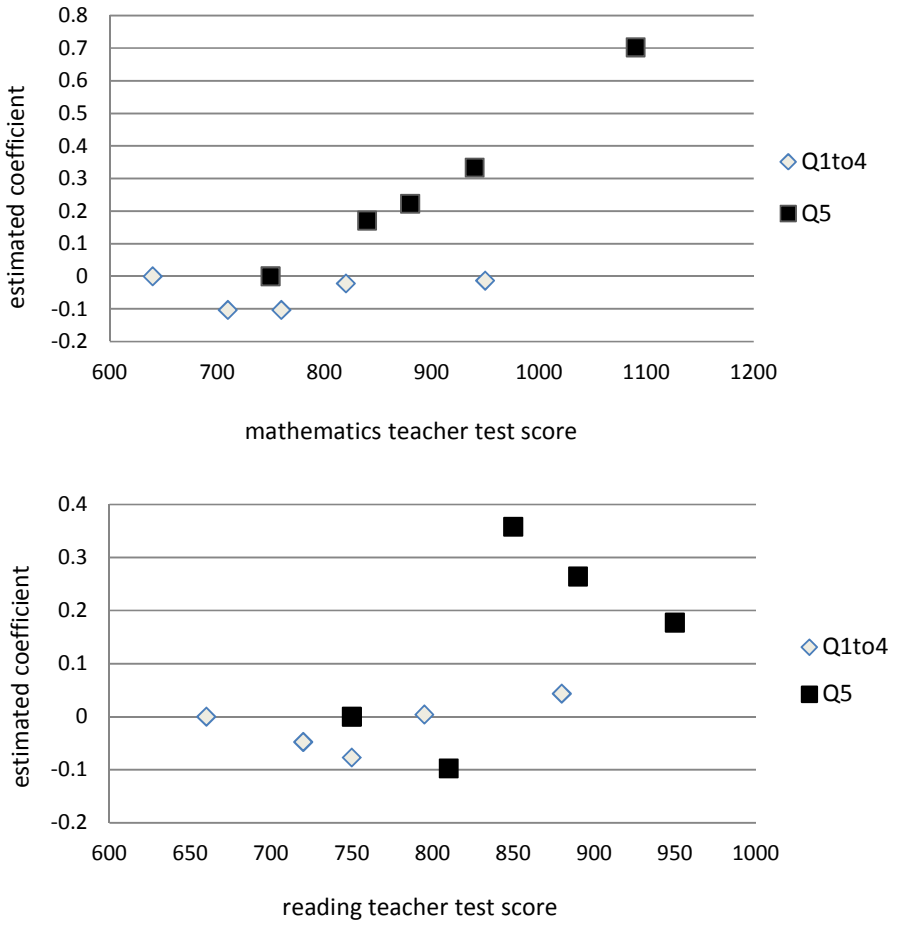
	(1)		(2)		(3)		(4)
	20% wealthiest schools				**80% poorest schools**		
	$\beta_1 \neq \beta_2$, $\eta_1 \neq \eta_2$		$\beta_1 \neq \beta_2$, $\eta_1 = \eta_2$		$\beta_1 \neq \beta_2$, $\eta_1 \neq \eta_2$		$\beta_1 = \beta_2$, $\eta_1 = \eta_2$
	Math	Read	Math	Read	Math	Read	Math/ Read
Implied β_s	0.110**	-0.042	0.115**	-0.050	-0.028	-0.006	-0.019
χ^2 $(\beta_s = 0)$	4.91	0.52	5.43	0.77	1.29	0.05	0.82
Prob > χ^2	0.027	0.471	0.020	0.379	0.256	0.823	0.366
Regression estimates:							
Teacher test score	0.177***	-0.087	0.130***	-0.035	0.070**	0.040	0.054**
in same subject	(0.068)	(0.075)	(0.048)	(0.060)	(0.035)	(0.034)	(0.022)
Teacher test score	-0.045	0.067	0.015		0.046	0.098***	0.072***
in other subject	(0.065)	(0.069)	(0.040)		(0.028)	(0.037)	(0.021)
$\chi^2(\eta_1 = \eta_2)$	1.07				0.01		
Prob > χ^2	0.301				0.908		
$\chi^2(\beta_1 = \beta_2)$	6.22**		8.61***		0.14		
Prob > χ^2	0.013		0.003		0.709		
Observations (students)	1317				5679		
Classrooms (clusters)	163				523		
Number of schools	65				260		

Notes: Dependent variable is the standardized (school sub-sample) student test score in numeracy and literacy. Regressions are estimated using seemingly unrelated regressions (SUR). Implied β_s is calculated as the difference in the coefficient on teacher test score in subject s between the equation of the student test score in the respective subject and the equation of the student test score in the other subject. In all models, the coefficients on student and school characteristics are constrained, with $\delta_1 = \delta_2$ and $\phi_1 = \phi_2$. Clustered standard errors, shown in parentheses and clustered at the classroom level, are estimated by maximum likelihood. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2 of the appendix to this chapter. Significance at *** 1% level ** 5% level * 10% level.

It is clear that there are great discrepancies in the role that teacher subject knowledge plays across the poorer and wealthier school sub-systems. Even in cases where teachers in Q1to4 schools possess high levels of subject knowledge that are comparable to that of teachers in Q5 schools, this is

not realized in the form of student performance gains. It should be acknowledged that the estimates on teacher knowledge in the Q5 sample may be upwardly biased by a correlation with unobservable teacher quality. Similarly, we may question whether or not the results for the group of Q1to4 schools may be driven by a *negative* correlation with teacher unobservables. Closer inspection of the data reveals that the test score variation of students taught by the least knowledgeable mathematics and reading teachers (scoring below 600 points) is the smallest. This might be indicative of effective teaching if it is believed that good teachers produce more equitable test outcomes. For example, a highly dedicated and enthusiastic teacher may not necessarily be the most knowledgeable teacher in terms of subject content, but he/she may more effectively transfer the knowledge they do possess, albeit small, to students. It could also be hypothesised that the working environment of teachers with adequate subject knowledge may be such that the benefits to teacher quality are not able to be realized.

Figure 5.3: Returns to teacher knowledge by performance quintile and school wealth group



Notes: the estimated coefficients are plotted at quintiles of the subject specific teacher test scores

Taylor and Taylor (2013) differentiate between three patterns of teacher knowledge in the SACMEQ III data, loosely named transmission, knowledge impedance and complex impedance. Transmission identifies those items in the test that both teachers and their students scored well on; hence teachers may well be affecting learning in these knowledge areas. Conversely, knowledge impedance and complex impedance patterns identify cases where teachers found it difficult to transmit knowledge, the first being due to a lack of knowledge on the part of teachers and the second due to an inability to convey knowledge. Correction for teacher unobservables will be explored in section 5.5.4.

5.5.3 Returns to teacher and classroom characteristics

Table 5.5 presents the estimated returns to other teacher and classroom characteristics aside from teacher content knowledge. Students attending a Q5 school taught by math and reading teachers with a university degree or post matric (diploma) qualification perform
approximately 20 to 40 percent of a standard deviation higher compared to students taught by teachers with less than higher education. A smaller positive effect of math teacher university education (11% of a standard deviation) is estimated for the sample of Q1to4 schools.

Surprisingly, a negative and statistically significant coefficient is estimated for diploma qualification of reading teachers in Q1to4 schools. Summary statistics indicate that reading teachers employed within Q1to4 schools with post-matriculation diplomas are older (significantly so) and more experienced than teachers with higher qualifications. It is likely that these teachers were trained under the former colleges of education that offered mainly diploma courses and have, since 1996, been absorbed into universities and other tertiary education institutions such as technical colleges. The majority of the students attending these colleges would not have obtained a matriculation exemption which would have allowed them access to a university degree. Many of the colleges were described as “glorified high schools” seen to be largely “underperforming and problematic in terms of turning out quality teachers” (Chisholm, 2009). Obviously this explanation for the negative diploma coefficient is conjecture. Clotfelter et al (2010) similarly find a negative effect size for teachers who invest in a postgraduate degree later into their teaching. This may be related to the recent provision of teacher qualification upgrades through the Advanced Certificate in Education (ACE). Unfortunately, the data does not provide information regarding the timing of receiving the diploma; it is therefore impossible to separate the causal effect of getting a diploma from the selection effect of the decision to get one.

The return to mathematics teacher experience is estimated to be 0.56 and 0.31 standard deviations for Q1to4 teachers with less than 5 years of experience and 6 to 15 years of experience, respectively. Similarly large effect sizes are found for mathematics teachers in Q5 schools, although they are less precisely estimated (possibly due to small sample sizes). The finding that the effect of teacher experience is highest in the first five years of teaching is in keeping with other research (Clotfelter et al, 2006, 2007) and may reflect the relative high quality of mathematics teachers who have recently entered the teaching profession following completion of formal training. Another interpretation is that very effective young, and therefore less experienced, teachers may opt out of teaching in government schools. The estimated coefficients on reading teacher experience are not estimated to be significantly different from zero for both school groups.

One of the most significant findings is the large positive and statistically significant effect of textbook availability on student achievement in poorer schools. Students having access to their own or a shared reading textbook has an estimated effect of 22 to 29 percent of a standard deviation increase in achievement, more than twice the effect size of being taught by a mathematics teacher with a university degree. Similarly, similarly high student access to mathematics textbooks is expected to increase math performance by 12 to 15 percent of a standard deviation in Q1to4 schools. This stresses the importance of adequate access to learning resources and teaching aids in South African classrooms, particularly for those students who are from disadvantaged socio-economic backgrounds.

5.5.4 Correction for teacher unobservables

When we compare the results of table 5.5 to the estimated teacher knowledge effects discussed in section 5.5.2, it is immediately evident that the estimated effect of teacher subject knowledge for the sample of Q1to4 schools is substantially smaller than that of other observable teacher and classroom characteristics. However, the estimates on teacher (and classroom) characteristics may be biased due to a correlation with the $(\tau_{1j_1} - \tau_{2j_2})$ component of the error term. For example, the large effect sizes of teacher qualification and teacher experience, as well as teacher subject knowledge in the Q5 sample, may be related to the quality of education and training received by teachers as was suggested by the findings of Carnoy and Chisholm (2008).

Table 5.5: Returns to other teacher and classroom characteristics

	20% wealthiest schools		80% poorest schools	
	Implied coef.	Prob > χ^2	Implied coef.	Prob > χ^2
	(1)		(2)	
Math teacher university degree	0.393***	0.001	0.100**	0.017
Reading teacher university degree	0.339***	0.002	0.005	0.900
Math teacher post-matric diploma	0.232*	0.082	0.010	0.861
Reading teacher post-matric diploma	0.203	0.166	-0.161***	0.002
Math teacher <5 years teaching experience	0.450	0.157	0.252**	0.045
Reading teacher < 5 years teaching experience	0.124	0.593	0.016	0.882
Math teacher 6-15 years teaching experience	0.220	0.462	0.117	0.337
Reading teacher 6-15 years teaching experience	-0.038	0.847	-0.080	0.460
Textbook shared between 2 students in math class	-0.139	0.183	0.119*	0.062
Students have their own textbooks in math class	-0.088	0.259	0.149**	0.019
Textbook shared between 2 students in reading class	0.127	0.283	0.289**	0.027
Students have their own textbooks in reading class	0.107	0.253	0.220*	0.087
Observations	1317		5679	
Clusters	163		523	
Schools	65		260	

Notes: Dependent variable is the standardized (sub-sample) student test score in numeracy and literacy. Regressions are estimated using seemingly unrelated regressions (SUR). Implied coefficients are calculated as the difference in the coefficient on the respective variable in subject s from the equation of the student test score in subject s and the equation of the student test score in the other subject. In all models, the coefficients on student and school characteristics are constrained, with $\delta_1 = \delta_2$ and $\phi_1 = \phi_2$. Clustered standard errors in the SUR models are estimated by maximum likelihood. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2 of the appendix to this chapter. Significance at *** 1% level ** 5% level * 10% level.

In order to correct for bias related to teacher unobservables, one can control for teacher fixed effects through restricting the analysis to the group of students who are taught by the same teacher for both subjects. The size of the same-teacher (referred to from this point onwards as ST) sample comprises of only 15 percent of the original student sample, which raises concern about the randomness of this sample. Inspection of the data reveals that schools within the ST sample are comprised of mostly rural, relatively poorer and smaller schools on the one hand (Q1to4), and relatively wealthier, urban and well-resourced schools on the other (Q5). This suggests that poorer schools in which teachers are observed to teach both subjects may do so out of necessity or lack of resources, whilst the opposite may be true of the wealthier school system that is able to attract highly educated teachers who are trained to teach several different subjects.

Comparisons of the student and teacher test score distributions of the Q5 ST sample to the Q5 non-ST sample reveals significantly higher performance in the former. In the case of Q1to4 schools, the ST sample of students and teachers performs significantly lower than the non-ST sample. In addition, students in the Q5 ST sample are significantly more likely to come from English speaking homes with more educated parents (particularly fathers) and more likely to be taught by younger, less experienced and more qualified teachers (all of which have been shown to have large positive effect sizes) than students within the Q5 non-ST sample. Conversely, students within the Q1to4 ST sample are significantly more likely to come from poorer homes with less educated parents and are taught in less resourced classrooms than the Q1to4 non-ST sample. However, teachers within the former sample are more likely to possess a university degree and spend significantly more time preparing for class (self-reported).

The results of estimating equations [5.6] and [5.7] are shown in table 5.6. The estimated effect sizes on teacher knowledge should be free from bias driven by teacher unobservables, at least subject-invariant ones. This, however, comes at the cost of lower precision given the smaller sample sizes. In both school ST samples we were not able to reject the over-identification restrictions and the final model was estimated with restrictions $\beta_1 = \beta_2$ and $\eta_1 = \eta_2$. The estimates for the ST sample of Q5 schools indicate an effect size of 5.4 percent of a standard deviation increase in student performance for a one standard deviation above average teacher subject knowledge, which is half that estimated for the whole sample of Q5 schools. A statistically significant effect size of teacher knowledge of 0.13 is estimated for the ST sample of Q1to4 schools. Whilst statistically insignificant, these effect sizes are in no way trivial.

The larger positive effect of teacher test score estimated for the Q1to4 schools when moving to the ST sample is suggestive of negative correlation between teacher subject knowledge and teacher unobservable characteristics. This is not to say that lower quality teachers necessarily perform better on the teacher test. Given that we know this group to be a relatively poorer subset of the whole Q1to4 sample, and hence also the overall South African school sample, we can expect the working environment to be such that the transmission of teacher knowledge to students may be hindered by a lack of teacher capacity; this may be linked on the one hand to poor formal training and a lack of strongly developed pedagogical skills, and on the other factors such as poor school leadership, overcrowded classrooms, absence of a learning culture and lack of community involvement (Bush, Joubert, Kiggundu & van Rooyen, 2010). If we further consider that the presence of the aforementioned factors are expected to be less prevalent (if not absent) in the Q5 ST sample that is likely to be

representative of the wealthiest and best performing schools, then it stands to reason that the smaller positive coefficient on teacher knowledge is indicative of a positive correlation between teacher knowledge and teacher quality unobservables.

Table 5.6: Correlated random error model results using the ST sample

	20% wealthiest schools (1) $\beta_1 = \beta_2, \eta_1 = \eta_2$		80% poorest schools (2) $\beta_1 = \beta_2,\ \eta_1 = \eta_2$	
	Maths	Reading	Maths	Reading
Implied β_s	0.054		0.130	
$\chi^2\ (\beta_s = 0)$	1.06		1.73	
Prob > χ^2	0.303		0.188	
Regression estimates:				
Teacher test score in same subject	0.109***		0.303**	
	(0.041)		(0.215)	
Teacher test score in other subject	0.055		0.173***	
	(0.043)		(0.187)	
Observations (students)	225		622	
Classrooms (clusters)	25		34	
Number of schools	14		32	

Notes: Dependent variable is the standardized student test score in numeracy and literacy calculated using the mean and standard deviation of the respective sample. Regressions are estimated using seemingly unrelated regressions (SUR). Implied coefficients are calculated as the difference in the coefficient on the respective variable in subject s from the equation of the student test score in subject s and the equation of the student test score in the other subject. In all models, the coefficients on student and school characteristics are constrained, with $\delta_1 = \delta_2$ and $\phi_1 = \phi_2$. Clustered standard errors (shown in parentheses) in the SUR models are estimated by maximum likelihood. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2 of the appendix to this chapter. Significance at *** 1% level ** 5% level * 10% level.

5.5.5 Fixed effects estimation

A number of the correlated random errors models estimated by this study have indicated that the over-identification restrictions are not rejectable. Does this then prescribe the use of a fixed effects model? The author would argue, not necessarily. The use of students as their own controls (as in the case of a fixed effects model) requires adequate within-student variability in the teacher and classroom characteristic. If variability is low (often referred to as sluggish covariates) then fixed effects estimation will lead to a fair amount of the share of variance in exposure to teacher content knowledge being removed and inflated standard errors. Both fixed effects and correlated random errors models are able to eliminate the bias in parameter estimates stemming from endogenous unobserved effects. As

mentioned it is difficult to argue that the error term $\tau_j + \varepsilon'_{si}$ will not contain some unobservable characteristics that are correlated with inter alia teacher subject knowledge, therefore we can expect some bias in the estimates regardless of estimation strategy chosen.[119] If, however, our intention is to estimate the effect of subject-invariant observable characteristics rather than to only control for them, then correlated random error modelling is the appropriate method.

In order to assess the appropriateness of the methodological strategy adopted by this study the estimates from the correlated random error models are contrasted with those from student fixed effects estimation; these are summarised in table 5.7. Despite being slightly larger, the model parameters on teacher subject knowledge are in general robust to those estimated using correlated random errors. It is expected that the coefficient on teacher knowledge for the sample of ST Q1to4 schools would be estimated with smaller standard error when fixed effect estimation is used. Sample-specific descriptives on the between- and within-student variation in teacher subject knowledge for the same samples considered in table 5.7 are presented in table 5.8. The within-student variation in teacher subject knowledge increases when the whole sample is sub-divided into the two school wealth groups. However, limiting the school wealth samples to those students taught by the same teacher in both subjects reduces the within-student variation in teacher knowledge. Although student fixed effect estimation appears to be a fair choice of methodological approach, and indeed provides results that are similar to that of a correlated random errors model, it is the opinion of the author that the latter approach is more adaptable when interest lies in estimating divergent effect sizes of teacher quality characteristics across different subjects.

5.6 Conclusion

In the South African context, where the vast majority of students perform at a level that is subpar both internationally and regionally, it is vitally important that we begin to understand the role that teachers play in schooling outcomes, and what the characteristics of high quality teachers are. Similarly, a better understanding is needed of the policy levers that will not only raise teacher quality in general, but also create a more equitable distribution of high quality teachers across the education system (Clotfelter et al, 2008: 3). The aim of this study was to add to the debate of the determinants of student performance in South Africa through identifying the impact of teacher content knowledge and other teacher and

[119] Fixed effects estimation assumes omitted variables to have time-invariant, or in this case subject-invariant, values as well as subject-invariant effects.

classroom factors on grade 6 student performance in reading and mathematics. To this end, the 2007 SACMEQ dataset and correlated random effects model estimation were employed.

Table 5.7: Student fixed effects estimation results

	Whole sample	20% wealthiest schools		80% poorest schools	
		All	ST	All	ST
Teacher test score	0.019	0.085**	0.063	-0.0002	0.152**
	(0.015)	(0.037)	(0.053)	(0.022)	(0.065)
Adjusted R-squared	0.020	0.091	0.039	0.021	0.025
Observations (students)	6996	1317		5679	
Classrooms (clusters)	686	163		523	
Number of schools	325	65		260	

Notes: Dependent variable is the standardized student test score in numeracy and literacy calculated using the mean and standard deviation of the respective sample. Robust standard errors clustered at the classroom level are shown in parentheses. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2 of the appendix to this chapter. Significance at *** 1% level ** 5% level * 10% level.

A number of important empirical findings emerge from this study and are discussed in turn. First, it is vital when estimating the impact of teacher and classroom factors on student outcomes that we control for unobservable school and student characteristics, as in the absence of these controls positive selection biases are observed on the estimates of teacher content knowledge. Accounting for selection biases on these unobservables, teacher knowledge is estimated to have no significant effect on student outcomes. This is similar to the findings of Carnoy and Chisholm (2008) and Carnoy and Arends (2012) who find no significant effect of teacher content knowledge on student gains in mathematics. However, this may mask differences in impact across student sub-groups.

This leads into the second important empirical finding that the impact of teacher knowledge is not homogenous across the South African education system. High quality teachers are typically observed to teach in Independent and former white and Indian schools that are likely to fall within the top school wealth quintile (Carnoy & Chisholm, 2008). Using average school SES as a proxy for former department and school wealth quintile, significant positive non-linear effects of teacher subject knowledge is estimated for the wealthiest quintile of schools. However, no significant effect of teacher knowledge is estimated for the poorest four school wealth quintiles. Teacher qualifications are estimated to have significant and large effects for student outcomes in wealthier schools, though this may be driven by a positive relationship to teacher unobservables. The same may be true of the large and highly significant effect size of young and inexperienced teachers in poor schools, which may signal an improvement in the training of those that have most recently entered the teaching profession.

Restricting the analysis to those students who are taught by the same teacher in both subjects removes any bias driven by a relationship between teacher unobservables and measurable teacher characteristics. Whilst the results for this sample may not be generalizable to the school system as a whole, they are likely to represent the two extremes of the South African education system; that is, the wealthiest of the Q5 schools and the poorest of the Q1to4 schools. The results indicate a positive effect size of teacher knowledge on performance of approximately 13-15 percent of a standard deviation and 5-6 percent of a standard deviation for the poorer subset and wealthier subset of South African schools, respectively. These estimates are in line with international findings that adopt similar techniques for estimating teacher effects. The most comparable of these studies is that of Metzler and Woessman (2012) who adopt an identical approach to that of this study in their assessment of the effect of teacher knowledge on grade 6 performance in Peru.[120] Metzler and Woessman's (2012) estimated effect size of 0.10 is very similar to that estimated for Q1to4 schools, as is that of Tan et al (1997) who find an estimated effect of teacher test scores of 0.10-0.12 on first grade learning gains in the Philippines. This illustrates that the findings for Q1to4 schools are largely in line with those of other developing country estimates. Conversely, the estimated effect size of teacher knowledge in Q5 schools is more comparable to the estimates found in developed country contexts, particularly the United States where estimates range between 0.01 and 0.06 (Hill et al, 2005; Goldhaber, 2007; Clotfelter et al , 2007).

The relationship between teacher knowledge and teacher unobservables further needs to be acknowledged. The analysis of this study suggests that teacher knowledge is positively related to teacher unobservable quality in Q5 schools, which we would expect. On the other hand, teacher knowledge appears to be negatively correlated to teacher (and school) unobservables in the poorest schools. This may be due to a lack of factors contributing to effective teaching such as high quality training, pedagogical skill and opportunity to teach that are more present in wealthier schools. It may also suggest a correlation with factors that hinder the transmission of knowledge to students such as mismanagement, poor instructional leadership and poor teacher collaboration. Clearly, not all teachers with poor content knowledge are ineffective teachers, and not all teachers with good content knowledge are effective teachers.

[120] A number of similarities can be drawn between South Africa and Peru. For example, the average performance of Peruvian students on international achievement tests also tends to be dismal when compared to developed countries. Furthermore, similar to the ranking of South African grade 6 students in SACMEQ III, Peruvian 6th grade students ranked 9 and 10 in mathematics and reading, respectively, amongst a comparative study of 16 Latin American countries from the Latin American Laboratory for Assessment of the Quality of Education (LLECE) in 2008.

A number of important policy conclusions arise from this study. First, the provision of textbooks and other teaching aides in poor schools is of utmost importance given the consistent finding by this study that the availability of textbooks to all students is associated with a large positive effect on performance. Furthermore, the effect size on textbook provision outweighs that of all other observable teacher and classroom characteristics identified in this study. The finding that the estimated effect size of teacher knowledge is of twice the magnitude in the poorest subset of schools reflects the relative importance of teacher knowledge for learning across the school system. Circumstance, both in the background of the teacher and the immediate working environment, will however dictate whether or not the benefits to teacher knowledge are able to be fully realized. The author would agree with Carnoy and Chisholm (2008) that the quality of teacher training and adequate curriculum preparation are crucial for explaining differences in student performance. Furthermore, the systematic differences with which high quality teachers are distributed across schools need to be addressed, if we consider this to be a driving factor behind the large performance gaps observed across school-wealth quintiles. School hiring practices need to take into account the long-term investment involved when selecting teachers, given their near-permanent employment statuses.

Appendix to Chapter 5

Table A5.1: Descriptive statistics (weighted) of selected variables (full sample)

Variable	Variable type	Mean	Standard deviation	Min.	Max.	Test score if indicator = 1[a]	
						Student	Teacher
Student test score							
Unstandardised:							
Numeracy	continuous	490.2	93.4	10.3	962.9		
Literacy	continuous	489.3	112.4	62.9	996.5		
Standardised:							
Numeracy	continuous	0	1	-5.153	5.017		
Literacy	continuous	0	1	-3.853	4.518		
Difference		0	1.392	-6.279	6.048		
Teacher test score							
Numeracy		0	1	-1.980	3.976		
Literacy		0	1	-2.607	4.122		
Student/family characteristics							
Female	dummy variable	0.506	0.500	0	1	0.074	0.014
Overage	dummy variable	0.436	0.496	0	1	-0.373	-0.202
Underage	dummy variable	0.088	0.283	0	1	-0.064	-0.100
Speak English most/all of the time	dummy variable	0.146	0.353	0	1	0.618	0.548
Never repeated	dummy variable	0.721	0.448	0	1	0.167	0.075
Repeated once	dummy variable	0.199	0.400	0	1	0.780	0.943
Repeated twice	dummy variable	0.052	0.222	0	1	-0.609	-0.253
Repeated > twice	dummy variable	0.028	0.164	0	1	-0.605	-0.308
Homework everyday	dummy variable	0.547	0.498	0	1	0.174	0.164
Homework 1-2 times/week	dummy variable	0.323	0.468	0	1	-0.124	-0.202
More than 10 books at home	dummy variable	0.279	0.449	0	1	0.530	0.393
Index of household chores	continuous	0	1	-1.773	3.446	-0.307	-0.240
Household socio-economic status*	continuous	0	1	-2.206	2.450	0.383	0.306
Mother has a matric qualification	dummy variable	0.174	0.379	0	1	0.176	0.188
Father has a matric qualification	dummy variable	0.220	0.415	0	1	0.074	0.092
Mother has higher level diploma	dummy variable	0.137	0.344	0	1	0.454	0.293
Father has higher level diploma	dummy variable	0.154	0.361	0	1	0.351	0.238
Mother has tertiary education	dummy variable	0.092	0.289	0	1	0.880	0.595
Father has tertiary education	dummy variable	0.118	0.322	0	1	0.659	0.467
Parents help with homework sometimes	dummy variable	0.567	0.496	0	1	0.129	0.083
Parents help with homework most of the time	dummy variable	0.345	0.475	0	1	-0.154	-0.116
School characteristics:							
School located in a town	dummy variable	0.181	0.385	0	1	0.146	0.028
School located in a city	dummy variable	0.293	0.455	0	1	0.600	0.521
School has a moderate absenteeism problem	dummy variable	0.327	0.469	0	1	-0.243	-0.072

Table A5.1: Descriptive statistics (weighted) of selected variables (full sample)

Variable	Variable type	Mean	Standard deviation	Min.	Max.	Test score if indicator = 1[a]	
						Student	Teacher
School resource index	continuous	0	1	-2.083	1.579	0.488	0.420
Lack of community involvement a problem	dummy variable	0.328	0.470	0	1	-0.109	-0.178
School average socio-economic status	continuous	0	1	-2.512	2.654	0.577	0.500
Classroom and teacher characteristics							
Only the teacher has a textbook	dummy variable	0.119	0.324	0	1	-0.128	-0.033
Textbook shared between > 2	dummy variable	0.142	0.349	0	1	-0.394	-0.232
Textbook shared between 2	dummy variable	0.264	0.441	0	1	-0.023	-0.059
Learners have their own textbook	dummy variable	0.394	0.489	0	1	0.250	0.158
Writing space:student ratio<1	dummy variable	0.704	0.457	0	1	-0.160	-0.167
Testing a few times term	dummy variable	0.467	0.499	0	1	-0.005	0.032
Testing done 2-3 times a month	dummy variable	0.240	0.427	0	1	-0.078	-0.150
Testing done weekly	dummy variable	0.142	0.349	0	1	0.204	0.133
Teacher female	dummy variable	0.611	0.488	0	1	0.037	-0.001
Teacher younger than 30 years	dummy variable	0.038	0.190	0	1	0.664	0.657
Teacher 31-40 years	dummy variable	0.438	0.496	0	1	-0.072	0.004
Teacher 41-50 years	dummy variable	0.372	0.483	0	1	-0.094	-0.084
Teacher has university degree	dummy variable	0.438	0.496	0	1	0.143	0.193
Teacher has a postmatric diploma	dummy variable	0.166	0.372	0	1	0.048	0.182
Teacher has 0-5 years' experience	dummy variable	0.119	0.324	0	1	0.021	-0.190
Teacher has 6-15 years' experience	dummy variable	0.386	0.487	0	1	-0.007	0.079
Teacher has 16-25 years' experience	dummy variable	0.421	0.494	0	1	-0.046	-0.026
Numbers of hours spent on preparation/week	continuous	10.117	7.669	0	25	0.080	-0.004
Number of in-service courses competed in last 3 years	continuous	3.533	5.121	0	61	0.075	0.014
Teaching minutes per week	continuous	1138.9	528.1	0	3000	0.174	0.177
Days lost due to strike activity	continuous	12.473	8.536	0	31	-0.323	-0.261

[a] For continuous variables these are mean standardised test scores for cases that are above the average, as given by the mean value of the continuous variable.

Notes: Household SES generated using principal component analysis on household possession items and standardized to have a mean of 0 and a standard deviation of 1; average school SES calculated as average of household SES within each school and standardized to have a mean of 0 and a standard deviation of 1.

Table A5.2: Classroom and teacher variables by subject

	Numeracy		Literacy		
Variable	Mean	Std dev	Mean	Std dev	Difference
Only the teacher has a textbook	0.162	0.368	0.059	0.236	0.102***
Textbook shared between > 2 learners	0.120	0.326	0.158	0.365	0.037***
Textbook shared between 2 learners	0.243	0.429	0.288	0.453	0.046***
Learners have their own textbook	0.366	0.482	0.442	0.497	0.076***
Writing space to learner ratio less than 1	0.668	0.471	0.684	0.465	0.016**
Class testing once a term	0.470	0.499	0.455	0.498	-0.015*
Class testing done 2-3 times a month	0.232	0.422	0.239	0.426	0.007
Class testing done weekly	0.156	0.363	0.148	0.355	-0.009
Teacher female	0.513	0.500	0.672	0.470	0.158***
Teacher younger than 30 years	0.047	0.212	0.037	0.188	-0.010***
Teacher 31 to 40 years	0.414	0.493	0.411	0.492	-0.003
Teacher 41 to 50 years	0.382	0.486	0.380	0.485	-0.003
Teacher has university degree	0.429	0.495	0.447	0.497	0.018**
Teacher has a postmatric diploma	0.178	0.383	0.160	0.367	-0.018***
Teacher has 0-5 years teaching experience	0.122	0.327	0.113	0.316	-0.009*
Teacher has 6-15 years teaching experience	0.363	0.481	0.374	0.484	0.011
Teacher has 16-25 years teaching experience	0.446	0.497	0.432	0.495	-0.014*
Numbers of hours spent on preparation/week	10.019	7.617	10.272	7.778	0.253*
Number of in-service courses competed in last 3 years	3.657	4.699	4.308	6.384	0.652***
Teaching minutes per week	1160.70	529.56	1218.68	525.30	57.98***
Days lost due to strike activity	12.110	8.462	11.868	8.648	-0.243*

Notes: significance at *** 1%, ** 5%, * 10%.

Table A5.3: Non-linear effects of teacher subject knowledge across school sub-systems

	20% wealthiest schools $\beta_1 \neq \beta_2, \eta_1 \neq \eta_2$ (1)		**80% poorest schools** $\beta_1 = \beta_2, \eta_1 = \eta_2$ (2)	
	Math	Reading	Math	Reading
Implied β_s (teacher score quintile 2)	0.209**	-0.085	-0.045	
Prob > χ^2	0.015	0.267	0.481	
Implied β_s (teacher score quintile 3)	0.322**		-0.073	
Prob > χ^2	0.010		0.213	
Implied β_s (teacher score quintile 4)	0.316***		0.008	
Prob > χ^2	0.001		0.886	
Implied β_s (teacher score quintile 5)	0.696***	0.225	0.046	
Prob > χ^2	0.000	0.139	0.522	
Observations	1317		5679	
Clusters	163		523	
Schools	65		260	

Notes: Dependent variable = standardized student test score in numeracy and literacy calculated using the mean and standard deviation of the respective school sub-sample. Teacher test scores are also normalized relative to the school sub-sample means and standard deviations. Regressions are estimated using seemingly unrelated regressions (SUR). Implied β_s is calculated as the difference in the coefficient on teacher test score in subject s between the equation of the student test score in the respective subject and the equation of the student test score in the other subject. In all models, the coefficients on student and school characteristics are constrained, with $\delta_1 = \delta_2$ and $\phi_1 = \phi_2$. Clustered standard errors (at the classroom level) in the SUR models are estimated by maximum likelihood. Regressions control for all student, classroom, teacher and school characteristics defined in tables A5.1 and A5.2. Significance at *** 1% level ** 5% level * 10% level.

Chapter 6

Compulsory tutorial programmes and performance in undergraduate microeconomics: A regression discontinuity design

(with Volker Schöer)[121]

As South African universities experience extremely low graduation rates, academic staff implement a range of interventions, such as tutorial programmes, in order to improve student performance. However, relatively little is known about the impact of such tutorial programmes on students' performance. Using data from an introductory microeconomics course, this paper investigates the impact of a compulsory tutorial programme on students' performance in their final examination. Due to the fact that the tutorial programme was only compulsory for students that obtained less than a pass in the first test, while otherwise offered on a voluntary basis, this paper employs a fuzzy regression discontinuity (RD) design to investigate the impact of the tutorial programme on final exam performance. Findings indicate that assignment to the compulsory programme positively affects students' performance. However, this result is mainly drive by students who already seem to have the ability to perform but, for whatever reason, underperformed in the first test. Thus, while assignment to the tutorial programme itself leads to an improvement in performance, the mechanism is unclear.

6.1 Introduction

At 15 percent, South Africa has one of the lowest university graduation rates in the world (Letseka & Maile, 2008). Drop-out rates amongst first-year students has also been reported to be as high as 35 percent at some universities during recent years. These worrying trends in higher education come at high financial and social costs. At the same time, university departments are taking strain as enrolment numbers continue to rise and resources are becoming even more limited. As a result of

[121] African Micro Economic Research Unit (AMERU), School of Economic and Business Sciences, University of Witwatersrand

these factors, university departments have the dual requirement of improving the quality of teaching while improving cost effectiveness ("doing more with less"). A further concern within the current teaching and learning environment of universities is that the traditional approaches to curricula and assessment have promoted a surface approach to learning rather than a deep or strategic approach which may bring disproportionate gains to minority student groups (Entwistle, Thompson & Tait, 1992). There are therefore both practical and ethical reasons for the move towards adopting peer tutoring as part of the learning support structure in higher education. The increase in use of peer tutoring in higher education courses clearly raises important questions of assessment, acceptance and the eventual success of such a programme, as poor design can be damaging to the positive features of what could be an important component of teaching and learning (Boud, Cohen & Sampson, 1999).

The specific microeconomics course used for purposes of this study initiated its own tutorial programme in 2009 that is run parallel to formal lecture sessions.[122] Attendance of these tutorials was made mandatory for poor performing students (obtained below 50 percent) who were identified through early assessment testing. Students who achieved at least 50 percent in the first test were still permitted to attend tutorials on a voluntary basis. The 2010 class cohort is used for analysis purposes given the stricter enforcement of the policy. The specific design of this policy has presented an opportunity to directly assess the impact of tutorial attendance on academic performance through the use of regression discontinuity design. Specifically, a fuzzy regression discontinuity design is employed to estimate a local average treatment effect of the tutorial programme within a bandwidth of the policy cut off. Estimates using both parametric and non-parametric models are presented.

This chapter begins with an overview of the literature that empirically investigates the effectiveness of peer tutoring on undergraduate performance in economics. The following section describes the data and policy design of the programme, followed by a discussion of the methodology. The next two sections present the empirical results and robustness checks, while the final section concludes.

6.2 Overview of the literature

The body of research on peer tutoring has seen tremendous growth in recent decades as illustrated by the many reviews and surveys (see Goldschmid & Goldschmid, 1976; Lee, 1987; Maxwell, 1989;

[122] The tutorial programme existed prior to 2009, although a full year undergraduate economics course was presented; that is, the first-semester microeconomics course was combined with the second-semester macroeconomics course. There is therefore limited comparability prior to and post 2009.

Frey & Whitman, 1990; Topping, 1996). The literature spans a range of elements of the peer tutoring process from practice to design and organisation (Schmidt & Moust, 1995), as well as assesses the relative advantages of peer tutoring for both tutees and tutors *inter alia* cognitive processes and emotional support as well as the impact on various outcomes such as performance, retention and drop-out. In determining the effectiveness of peer tutoring, one should be cognisant that programmes tend to be diverse and therefore may have very little in common. For example, tutors may be staff or students; the tutor and tutees may meet in individual or group settings; frequency of meetings may range from several times a week to once a week to once a month; tutors may receive special training or may be unsupervised; tutors may receive some form of remuneration or may volunteer to participate; tutors and tutees may have some or no choice in their pairings; and so forth. Additionally, tutoring programmes may differ in their aims and objectives, be it improved achievement, reduced attrition or increased interest in the subject. Three methods of peer tutoring have been widely used in higher education and have demonstrated to be quite effective (Topping, 1996). These are: cross-year small-group tutoring, where upper year undergraduates or postgraduates function as tutors to a small group of lower undergraduate students; the personalised system of instruction (PSI), where students are able to progress through the study material at their own pace and the role of peer tutors are largely to check, test and record the advancement of tutees; and supplemental instruction (SI).

The evaluation of peer tutoring programmes in higher education has traditionally tended to use weak programme designs, with much of the empirical work relying on cross-sections of subjective outcome measures that are largely retrospective in nature (Jacobi, 1991). Often the data are reported without adequate evidence of reliability and validity. However, recent research has become more empirically rigorous, with greater use of experimental and randomly controlled programme designs that attempt to correct for potential selection biases. While student-to-student tutoring has been used with some success in several disciplines, there have been relatively few evaluations of its impact on student learning in economics (see Kelley & Swartz, 1975; Munley, Garvey & McConnell, 2010). Research is even more limited in a South African context (Horn & Jansen, 2009).

The few empirical studies that have been published tend to be fraught with methodological weaknesses that seriously limit both internal and external validity of the results. For example, research of tutorial programmes that are based on systematic selection rather than random assignment need to make adequate attempts to control for sampling and self-selection biases, although there should be recognition that the corrections are likely to be imperfect or incomplete (Cook, Campbell & Peracchio, 1990). A further concern problem with peer tutoring research is the

potentially low levels of external validity. Most research is based on data collected within a single department within a specific university. The scope for generalizing these findings based on these studies to other tertiary institutions and other students is limited.

A study of a peer tutoring programme at Duke University by Kelley and Swartz (1975) made use of weekly computer based tests to differentiate between good and poor performing students after which the top performers were given the option to tutor weaker students in exchange for exemption form a forthcoming examination. The performance of students who accepted an invitation to attend the tutorial sessions was compared to the group of students who declined. A significant positive impact of 0.67 standard deviations (4.2 percentage points) on the final course score was estimated. However, it is posited that these results may understate the true impact of tutorials as it excludes the performance of the tutors themselves. The authors correctly recognise that the group of tutees are a self-selected group and that their results are likely to be inflated by selection on unobservables, most notably motivation, despite the two groups being very similar on observables.

In a South African context, Jansen and Horn (2009) make use of ordinary least squares regression to model the impact of various factors, including tutorial attendance, on the course mark in an undergraduate economics course at a South African university. Student attendance of these tutorials was voluntary, although students who performed poorly in the first test were encouraged to attend. The group of students who attended regularly were found to have better school-leaving grades and a better average performance in economics. This therefore raises concerns that the coefficient on tutorial attendance may be biased due to sample selection. Class attendance was included as a proxy for motivation, which may serve as a control against the voluntary attendance. A significant positive effect of tutorial attendance on performance was found, with a larger effect for first-time registered students than for repeat students.

More recently a number of studies have attempted to estimate the impact of peer tutor programs through experimental design so as to correct for selection bias. Johnston and James (2000) evaluate the impact of a collaborative, problem solving (CPS) approach to tutorials in a second-year macroeconomics course. Treatment and control groups were generated where one group was exposed to the CPS approach whilst the other attended tutorials that continued to use the traditional approach. Programme evaluation was based on qualitative measures such as student attitude and teaching-evaluation questionnaires, as well as quantitative information regarding tutorial attendance and examination performance. Students attending CPS were found to both value their tutors' performance and enjoy their tutorials more. They also spent significantly more time preparing for the tutorial sessions. No consistent gain was observed for the control versus

treatment groups, except in the case of foreign students. The researchers posit, though not convincingly, that the non-significant change in performance and learning may be due to spill over effects or inappropriate selection of the control and treatment groups.

Munley et al (2010) evaluate the effect of participating in a tutoring programme across several courses and several years (including undergraduate economics) using two methodological approaches. First they model the exogenous effect of participation or level of participation on the final grade; and second, given voluntary participation, they adopt a treatment model defined by Greene (2003) where participation and performance are modelled jointly using selection and outcome equations. They use two policies regarding intercollegiate athletics as an exclusion restriction. Under the first model treating participation as exogenous, they find a *negative* and statistically significant coefficient on the binary choice to participate in tutorials, which they put down to participation likely being higher amongst weaker students. Modelling the choice to participate, the coefficient on tutorial participation turns positive but is statistically insignificant. However, modelling the level of participation rather than the choice to participate yields positive and significant results. Therefore, the amount of participation appears to be more relevant for improving performance, with a sufficient amount of tutorial attendance required in order to see notable gains.

It is clear from the already existing research that the results are mixed, which may in part be due to differences in the underlying programmes and their participants, or the choice of modelling strategy. This study aims to add to the current empirical evidence on the effectiveness of peer tutoring in economics through the use of what the authors believe to be a truly exogenous tutorial programme that addresses the issue of sample selection bias.

6.3 Data and Policy Design

This study uses tutorial attendance data from an undergraduate micro economics course that was run during the first academic semester (February to May) of 2010 at Stellenbosch University. The course has one of the largest enrolments amongst undergraduate modules at the university, with 1767 students enrolled in the year analysed.[123] Students were sub-divided by language (English or Afrikaans) into one of seven formal lecture classes. Students were expected to attend three 50-minute lectures per week for 14 weeks, as well as one 50-minute tutorial session that began two weeks after the start of the formal academic semester and lasted for the remaining 12 weeks of the semester. The tutorial programme is one of structured academic support where students are able

[123] 27 students unenrolled themselves during the course of the semester, and are therefore dropped from the analysis.

to benefit from a small-class environment (less than 30 students per tutorial). Students are instructed to attempt a tutorial question set that tackles problems related to coursework material covered in the formal lectures in the preceding week. This is provided to all students one week prior to the tutorial. Tutors are expected to cover as many of the answers to these problem sets, time permitting.

Attendance of tutorial classes is voluntary up until a week following the first semester test, after which students with a test score below a passing score of 50 percent were required to attend the tutorial classes on a compulsory basis. Students who did not write the first semester test were also subject to compulsory tutorial attendance. Given that we do not observe their performance, and therefore cannot necessarily include them in the group of "just failers", these students are dropped for analysis purposes. Furthermore, in order to make comparisons from test 1 to test 2, we only consider those students who wrote both tests. Our final sample is therefore comprised of 1653 students (93.5 percent of the original sample). Tutorial attendance remained voluntary for students that scored at least or above 50 percent in the first test. The compulsory tutorial policy was announced in the first week of classes, with further reminders given in the weeks prior to and after the first test. Students that scored below 50 percent were alerted via e-mail that they were required to attend the tutorial classes. Tutorial attendance was recorded by tutors as students arrived for each tutorial class. Any student that left before the end of the tutorial was not marked down as attending.

The first semester test (or early assessment test) was written fairly early into the semester a few weeks after the start of tutorial classes.[124] In addition to this early assessment test, students are required to write at least one of two remaining semester tests, although students are permitted to write all three if they choose. Admission to the examination is contingent on achieving an average of at least 40 percent on the semester tests. Therefore, whilst 1767 students were enrolled for the course at the beginning of the academic year, only 1489 achieved the required semester average to gain entrance to the exam. Those students who did not gain access are likely to be compulsory tutorial students. This could cause concern for our analysis as the sample of students between test 2 and the exam are not the same. However, given that we are only interested in the effect of tutorial attendance for students who perform within a neighbourhood around the cut off, the two samples are unlikely to be that dissimilar. Students were also offered a choice of writing one of two exams, both of which are set to be of the same difficulty. Students who wrote the first exam and did not

[124] The first semester test comprised of 10 true/false and 10 multiple choice questions (referred to as short answer questions). Subsequent tests and exams consisted of both descriptive and short answer questions. Tests are marked by postgraduate teaching assistants, whilst the course lecturers are involved in the marking of the examinations. In general, markers are unaware of which students are subject to compulsory tutorial attendance.

achieve a passing mark (50 percent) but achieved a sub-minimum weighted average of 40 percent for their semester tests and exam were permitted to write the second examination option. Students who chose only to write the second exam therefore only received one exam opportunity. For purposes of this study, we consider the mark obtained by the student in their first exam attempt.[125] As part of the course administration, each student's tutorial attendance, tutoring sessions attended, semester test, class mark and final exam scores, gender, year of enrolment and degree major were recorded. Additional information regarding the student's high school leaving performance, school department, home language, age and test scores in additional undergraduate courses taken in the same semester were also obtained.

6.4 Methodology: the Regression Discontinuity Design

We are interested in estimating the effect that participation in the tutorial programme, T_i, has on test scores Y_i. We assume that Y_i is further related to some vector of observables W_i, such that:

$$Y_i = \beta_0 + \alpha T_i + W_i\beta_1 + u_i \quad [6.1]$$

where α represents the effect of T_i, assumed to be constant across individuals, and the error term ε_i is assumed to be uncorrelated with W_i. Unless treatment has been randomly assigned conditional on W_i, identification of α is hampered by selection bias due to some dependence between T_i and u_i. This arises when treatment is related to some unobservable/s not included in W_i. The resulting dependence between T_i and u_i will therefore be erroneously attributed to the impact of the programme on the outcome of interest.

We solve for the selection issue using information about the mechanism by which participation in the tutorial programme was assigned. Specifically, compulsory tutorial attendance was determined by performance in the first semester test: students scoring below a given cut off *c* (50 percent) were required to attend tutorials on a mandatory basis, while students scoring at or more than c were not subject to the compulsory tutorial policy. Therefore, students are assigned to tutorials based on the following deterministic rule:

$$D_i(X_i) = 1\{X_i \geq c\} \quad [6.2]$$

where X_i is student *i*'s first semester test score, *c* is the cut off test score and 1{.} is the indicator function.

[125] Robustness checks will be performed considering the final exam mark following all attempts, as well as controlling for whether or not the student chose to write the second exam option or not (if we believe that weaker students are more likely to delay sitting the exam). A comparison of means indicates that compulsory students are no less likely to write the second option than non-compulsory students are. However, compulsory tutorial students are more likely to write both exam options. This is to be expected given that they are weaker performing students.

The above corresponds to the selection rule of a sharp Regression Discontinuity design (Thistlethwaite & Campbell, 1960). The assignment mechanism is clearly not random (there is little reason to suppose that X_i is unrelated to Y_i), therefore a simple comparison of means between the treatment and non-treatment (control) groups would not suffice to provide an unbiased estimate of α. However, if we expect that for some arbitrarily small number $\varepsilon > 0$ that $E[\alpha_i|X_i = c + \varepsilon] \cong E[\alpha_i|X_i = c - \varepsilon]$ and further assume that both $E[u_i|X]$ and $E[\alpha|X]$ are continuous in X at c (Hahn, Todd & van der Klaauw, 2001; Van der Klaauw, 2002), then we have:

$$\lim_{\varepsilon \downarrow c} E[Y|X] - \lim_{\varepsilon \uparrow c} E[Y|X] = \alpha \qquad [6.3]$$

Therefore, by comparing individuals arbitrarily close to *c* who did and did not receive treatment, we are able to identify (in the limit) the causal impact of the tutorial programme on performance.

However, given that tutorials were not denied to the group of students scoring at or above the cut off, the rate of tutorial attendance as a function of semester test 1 performance is now a discontinuous function in X_i at *c*. This represents the discontinuity "fuzzy" or stochastic RD design. Under the same two continuity assumptions listed above and the additional assumptions of local "monotonicity"[126] and "excludability"[127] (Hahn et al, 2001; Imbens & Angrist, 1994) gives:

$$\frac{\lim_{\varepsilon \downarrow c} E[Y|X] - \lim_{\varepsilon \uparrow c} E[Y|X]}{\lim_{\varepsilon \downarrow c} E[T|X] - \lim_{\varepsilon \uparrow c} E[T|X]} = \alpha_F \qquad [6.4]$$

where the subscript *F* represents the fuzzy treatment estimator. Taking the limit of both sides of (4) as $\varepsilon \rightarrow c$ would identify the "local Wald" estimator, α, as in Hahn et al (2001):

$$\alpha_F = \frac{Y^+ - Y^-}{T^+ - T^-} \qquad [6.5]$$

6.4.1 Estimation

Parametric: IV estimator

In a context such as this where treatment is continuous (T) and there is a randomized binary instrument (D), an instrumental variable (IV) approach is an obvious way of obtaining an estimate of the impact of T on Y . The treatment effect, $\alpha_{IV,}$ is calculated as the reduced form impact D on Y divided by the first-stage impact of D on T, and uses the entire sample of observations. The model set-up is the same as in [6.1] except with an added second equation that allows for imperfect compliance and observables and unobservables to impact the rate of tutorial attendance:

[126] X crossing c cannot simultaneously cause some units to take up and others to reject.
[127] X crossing c cannot impact Y except through impacting receipt of the treatment.

$$Y_i = \beta_0 + \alpha T_i + W_i\beta_1 + \varepsilon_i$$
$$T_i = \theta_0 + D_i\pi + W_i\alpha + \nu_i$$
$$D_i = 1.\{X \geq c\}$$
$$X = \beta_2 W_i + \xi_i \quad [6.6]$$

where we make no assumptions about the correlations between W, ε, ν and ξ. It is simple to show that:

$$\lim_{\varepsilon \downarrow c} E[Y|X = c + \varepsilon] - \lim_{\varepsilon \uparrow c} E[Y|X = c + \varepsilon] = \{\lim_{\varepsilon \downarrow c} E[T|X = c + \varepsilon] - \lim_{\varepsilon \uparrow c} E[T|X = c + \varepsilon]\}\alpha \quad [6.7]$$

where the left-hand side represents the reduced-form discontinuity in the relation between Y and X, and the term in front of α is the "first-stage" discontinuity in the relation between T and X. The ratio of the two discontinuities yields the treatment estimator α.

There is no particular reason to believe that the true model is linear, and the consequences of incorrect functional form are more serious in the case of RD design as misspecification generates bias in the estimator of interest, α. Allowing for non-linearities in the underlying function of X can be important, especially in cases where we suspect X and Y to be non-linearly related, for example, when we have reason to expect this relationship to change as a result of the program. One way of circumventing this is to augment the outcome equation with a regression function $f(X)$, known as the control function approach (Heckman & Robb, 1985). We can generalise this function by allowing the X_i terms to differ on each side of the cut-off by including the X_i terms both individually and interacted with D_i (Van der Klaauw, 2002; Lee & Lemieux, 2010; McCrary, 2008). The reduced-form outcome function is now:

$$Y_i = \beta_0 + D_i\alpha\pi + \delta_{01}\tilde{X}_i + \delta_{02}\tilde{X}_i^2 + \cdots + \delta_{0p}\tilde{X}_i^p + \delta_1 D_i\tilde{X}_i + \delta_1 D_i\tilde{X}_i^2 + \cdots + \delta_p D_i\tilde{X}_i^p + W_i\beta_1 + \varepsilon_i \quad [6.8]$$

We can also allow for a control-function $g(X)$in the first-stage equation:

$$T_i = \theta_0 + D_i\pi + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p + \gamma_1 D_i\tilde{X}_i + \gamma_1 D_i\tilde{X}_i^2 + \cdots + \gamma_1 D_i\tilde{X}_i^2 + W_i\alpha + \nu_i \quad [6.9]$$

where $\tilde{X}_i = (X_i - c)$. The instrumental variable estimate of treatment is obtained by taking the ratio $\alpha\pi/\pi$. Given that the model is exactly identified, a two-stage estimation procedure per van der Klaauw (2002) will be numerically identical to $\alpha\pi/\pi$. This involves estimating the control function augmented second-stage outcome equation by replacing T_i with the first stage estimate. With

correctly specified control functions *f(X)* and *g(x)*, this two-stage procedure yields a consistent estimate of the treatment effect. If we assume the same functional form for *f(x)* and *g(x)*, then the two-stage estimation procedure described here will be equivalent to a two-stage least squares estimation with $\mathrm{D_i}$ and the terms in *f(x)* serving as instruments.

It should be noted that the instrumental variable estimate may still be biased by omitted variables if the compulsory tutorial policy changes student behaviour with regards to other learning such as studying, effort and class attendance. Student behaviour may be adjusted in a number of ways: first, students who are required to attend Economics tutorials on a mandatory basis may decrease the amount of time they spend studying or attending class, thereby underestimating the impact of the tutorials; secondly, compulsory tutorial students may feel that there is a stigma attached to the programme, and therefore will put in more effort than students who just passed semester test 1, leading to an overestimate in the impact of tutorials.

Non-parametric: Wald estimator

The estimation procedure described above is a parametric one that uses polynomial regression. Parametric estimation typically uses data away from the cut off, therefore providing global rather than local estimates of the regression function. However, in practice one can consider using a narrower window of observations around the cut off. Non-parametric techniques offer more flexible estimates of the regression function, as well as address the "boundary problem" of RD (we are interested in computing an effect at the cut off using only the closest observations). We could consider using kernel regression given that it is well suited from estimating regression functions at a particular point. However, in finite samples, precise estimation requires sufficiently wide bandwidths, and wider bandwidths come at the cost of greater bias. Local linear regressions have been introduced as a means of reducing bias in standard kernel regression methods (Fan and Gijbels, 1995; Hahn et al., 2001). Estimates under local linear regression are obtained by solving:

$$\min_{a,b} \sum 1(X_i \geq c)(y_i - a - b(X_i - c))^2 K\left(\frac{X_i - c}{h}\right) \qquad [6.10]$$

in the case of $Y^+ = \lim_{\varepsilon \downarrow c} E[Y|X = c + \varepsilon]$, and:

$$\min_{a,b} \sum 1(X_i < c)(X_i - a - b(X_i - c))^2 K\left(\frac{X_i - c}{h}\right) \qquad [6.11]$$

in the case of $Y^- = \lim_{\varepsilon \uparrow c} E[Y|X = c + \varepsilon]$, with K(.) a kernel function and h a bandwidth that converges to 0 as $n \to \infty$. Estimates for T^+ and T^- are found in a similar way.

Various techniques are available for choosing the kernel function and bandwidths. Less important is the choice of kernel. RD design studies tend to adopt either the rectangular or

triangular kernels, with the difference between the two that the latter places more weight on observations close to the cut off. Of more importance is the choice of bandwidth, as different bandwidth choices can produce quite different estimates. For this reason, it is sensible to report at least three estimates as an informal sensitivity test: one using the preferred bandwidth, one using twice the preferred bandwidth and another using half the preferred bandwidth (McCrary, 2008). In general, choosing a bandwidth in non-parametric estimation involves finding an optimal balance between precision and bias. The default bandwidth from Imbens and Kalyanaraman (2012) is designed to minimize MSE (squared bias plus variance) in a sharp RD design. However, the optimal bandwidth will tend to be larger for a fuzzy design due to the additional variance arising from the estimation of the jump in the conditional mean of treatment. Unfortunately, a larger bandwidth also leads to additional bias. According to McCrary (2008), the best method of bandwidth selection is visual inspection guided by an automatic procedure. A simple automatic bandwidth selection procedure uses a rule-of-thumb (ROT) bandwidth as follows:

$$h_{ROT} = \kappa \left[\frac{\tilde{\sigma}^2 R}{\sum_{i=1}^{N} \{\bar{m}''(X_i)\}^2} \right]^{1/5} \qquad [6.12]$$

where κ is 2.702 (3.348) in the case of the rectangular (triangular) kernels respectively, $\tilde{\sigma}^2$ is the estimated standard error of a 4th order polynomial regression of X on Y, R is the range of X and $\bar{m}''(X_i)$ is the second derivative implied by the global polynomial model (Fan & Gijbels, 1995). Imbens and Lemieux (2008) recommend using the same bandwidth in the treatment and outcome regressions. When we are close to a sharp RD design, *g(X)* is expected to be very flat and the optimal bandwidth to be very wide. In contrast, there is no particular reason to expect the *f(X)* to be flat or linear, which suggests the optimal bandwidth would likely be less than the one for the treatment equation. As a result, Imbens and Lemieux (2008) suggest focusing on the outcome equation for selecting bandwidth, and then using the same bandwidth for the treatment equation.

6.4.2 Inclusion of covariates

Up to this point estimation has explicitly allowed for the inclusion of baseline observables as covariates in the regression models. The baseline covariates are useful for testing the validity of the RD design by testing that the local continuity assumptions are satisfied. In their capacity as additional controls for parametric and non-parametric estimation, the only possible gain this affords is a reduction in the sampling variability (assuming they have explanatory power). However, estimation error in their covariates could also reduce efficiency. If the RD design is indeed valid, that is, the distribution of W given X is continuous at the threshold; the inclusion of additional covariates should still provide a consistent estimate of the local treatment effect (Imbens & Lemieux, 2008:

626). If including these controls leads to significant changes in estimates, this would suggest that the continuity assumptions may be violated and the treatment estimates are likely to be biased. Lee (2008) proposed a method to test the sensitivity of RD estimates to the inclusion of covariates by first regressing Y on a vector of individual characteristics and then to repeat the RD analysis using the residuals $\left(Y_i - \hat{Y}_i\right)$ as outcome variable. Intuitively, this procedure nets out the portion of the variation in Y we could have predicted using the pre-determined characteristics, making the question whether the treatment variable can explain the remaining residual variation in Y. The important thing to keep in mind is that if the RD design is valid, this procedure provides a consistent estimate of the same RD parameter of interest.

6.5 Results

From Figure A6.1 of the appendix to this chapter it is clear that prior to semester test 1 there was variation in the number of tutorials attended by students. Approximately 55 percent of students attended all tutorials, while more than a fifth of all students did not attend any of the voluntary tutorials. Figure A2 shows how weekly tutorial attendance changed over the semester by compulsory and non-compulsory status. Week 0 indicates the week in which semester test 1 was written. It is immediately clear that prior to test 1, attendance amongst the non-compulsory group was higher than that of the compulsory group. Attendance amongst both groups also appeared to drop during the week in which the first semester test was written. Once the mandatory policy was instituted, the attendance of the compulsory group is approximately 40 percent higher than the non-compulsory group. Noting the trend in tutorial attendance amongst the group of non-compulsory students, the mandatory policy appears to have worked to counteract the tendency for tutorial attendance to decline over the semester.

Table 6.1 compares the average characteristics of the group of compulsory tutorial students with those of the group of non-compulsory students. Observing the entire student sample, students in the compulsory group were significantly less likely to attend tutorials prior to writing test 1. Additionally, members of this group are more likely to be repeat students, registered for degrees other than Actuarial Science, Accounting, Law and Mathematics and have achieved a lower matric[128] maths mark. They are also less likely to have been part of the NSC[129] matriculant cohort. These differences suggest that identification of the impact of tutorial attendance on test and exam performance using OLS regression would very likely suffer from omitted variables bias. In terms of

[128] "Matric" refers to the final examination at the end of secondary schooling in South Africa. The matriculation exams are centrally set and standardized which allows comparisons of students' abilities that graduate from different secondary schools.
[129] The National Senior Certificate (NSC) is currently the school leaving certificate in South Africa. It replaced the Senior Certificate (SC) with effect from 2008 and was phased in starting with grade 10 in 2006

demographics, the only distinguishing feature of the two groups is that compulsory students are less likely to be from the white population group and more likely to be home-language Afrikaans speakers. The mandatory tutorial policy has a significant effect on attendance subsequent to semester test 1, with compulsory students attending 45 percent more tutorials prior to the exam when considering the entire sample.

When the sample is narrowed to within 1 and 0.5 standard deviations from the policy threshold, the tutorial attendance gap prior to test 1 turns insignificant. The differences in post policy performance are also reduced. Despite the substantial increase in tutorial attendance of compulsory students relative to non-compulsory students, the latter continue to significantly outperform the former in tests, despite attending fewer tutorials. However, there are no notable differences in exam performance once the window is narrowed to 0.5 standard deviations. This may be due to the fact that, even with the narrower window, we are still capturing students of differing abilities (note a significant difference in matric maths performance for this sample). The final column of table 6.1 displays coefficients on the binary treatment from a regression of each of the characteristics on the quadratic control function from equation [6.2] without any covariates. These estimates describe how each variable differs between the compulsory and non- compulsory groups at the policy threshold. It is evident that, at the threshold, the post-test 1 attendance rate is significantly higher for the group of compulsory students. The difference in test and exam performance across the policy threshold is negative and statistically significant (at the 5 and 10 percent levels). This indicates that, at least within a window around the cut-off, performance is higher for the group of compulsory students. There is no significant difference in the other outcomes or characteristics.[130]

Figures A6.3, A6.4 and A6.5 of the appendix present scatter plots of the average tutorial attendance prior to and after test 1 in 0.1 standard deviation wide bins of the normalised test 1 score. It is clear that there is no noticeable discontinuity in attendance prior to test 1 at the threshold. However, once the compulsory tutorial policy is instituted, there are clear discontinuities in attendance at the policy threshold prior to the second semester test and the exam, with non-compulsory student behaviour appearing to change very little between test 2 and the end of the semester. The figures further display linear, quadratic and cubic fits to the underlying data. A linear fit of the running variable appears to capture the primary relationship between attendance and test 1 scores the best. The primary analysis will therefore employ a linear form of the control function, with results based on alternative functional forms generated as robustness checks.

[130] Except for the Eastern Cape Education Department.

Table 6.1: Comparison of compulsory and non-compulsory tutorial attendance groups

	Whole sample		1 SD from threshold		0.5 SD from threshold		Parametric RD
	Non-comp	Comp	Non-comp	Comp	Non-comp	Comp	
percentage tutorials prior to test 1	0.7276	0.5927	0.6948	0.5973	0.6483	0.5851	0.0727
	0.1349***		0.0975***		0.0632		
percentage tutorials prior to test 2, post-test 1	0.5420	0.7325	0.5197	0.7368	0.4929	0.7207	-0.2305***
	-0.1905***		-0.2171***		-0.2278***		
percentage tutorials prior to exam, post -test 2	0.4317	0.8818	0.4218	0.8838	0.4056	0.8915	-0.4348***
	-0.4501***		-0.4620***		-0.4858***		
normalised test 2 score	0.6671	-0.2869	0.3107	-0.1371	0.1361	-0.0445	-0.3094***
	0.9540***		0.4478***		0.1807***		
normalised exam mark (first attempt)	0.4400	-0.3337	0.0968	-0.3029	-0.0752	-0.1725	-0.3608**
	0.7738***		0.3997***		0.0972		
Female	0.4441	0.4446	0.4281	0.4161	0.4310	0.3850	0.0270
	-0.0005		0.0120		0.0460		
Degree other	0.4800	0.7563	0.5757	0.7346	0.6158	0.7181	0.0861
	-0.2763***		-0.1588***		-0.1023**		
BA (PPE/VPS)	0.0439	0.0434	0.0457	0.0412	0.0508	0.0532	-0.0322
	0.0005		0.0046		-0.0023		
BAccounting	0.3149	0.1219	0.2524	0.1350	0.2175	0.1383	-0.0420
	0.1930***		0.1174***		0.0792**		
BComm (Actuarial Science)	0.0658	0.0033	0.0315	0.0046	0.0198	0.0053	0.0023
	0.0625***		0.0270***		0.0145		
BComm (law/maths)	0.0754	0.0618	0.0741	0.0664	0.0791	0.0585	0.0075
	0.0136		0.0078		0.0206		
BComm (Economics)	0.0200	0.0134	0.0205	0.0183	0.0169	0.0266	-0.0218
	0.0067		0.0022		-0.0096		
repeater	0.0620	0.1269	0.0804	0.1373	0.1073	0.1489	-0.0682
	-0.0649***		-0.0569***		-0.0416		

Table 6.1 continued: Comparison of compulsory and non-compulsory tutorial attendance groups

	Whole sample		1 SD from threshold		0.5 SD from threshold		Parametric
	Non-comp	Comp	Non-comp	Non-comp	Comp	Non-comp	RD
White	0.8613	0.7752	0.8307	0.7908	0.8247	0.7647	0.0971
	0.0861***		0.0399		0.0600		
Afrikaans	0.4875	0.5638	0.4649	0.5678	0.4540	0.5775	0.0151
	-0.0763***		-0.1030***		-0.1235**		
English	0.4123	0.3591	0.4313	0.3563	0.4339	0.3529	0.0136
	0.0533**		0.0750**		0.0810*		
Age	19.3015	19.3222	19.2939	19.3012	19.2730	19.3369	-0.0205
	-0.0206		-0.0072		-0.0639		
NSC	0.1183	0.1269	0.1309	0.1190	0.1412	0.1223	-0.0270
	-0.0086		0.0119		0.0189		
normalised matric maths score	1.9835	1.1501	1.7193	1.2307	1.6062	1.3338	-0.0699
	0.8333***		0.4886***		0.2724***		
normalised matric maths score * NSC	0.1126	0.0142	0.1164	0.0175	0.1271	0.0190	0.0844
	0.0983***		0.0990***		0.1081**		
Gauteng ED	0.0867	0.0807	0.0831	0.0737	0.0948	0.0538	0.0425
	0.0060		0.0093		0.0411		
OEB	0.1638	0.1294	0.1629	0.1475	0.1724	0.1290	0.0348
	0.0344		0.0155		0.0434		
Eastern Cape ED	0.0530	0.0454	0.0511	0.0507	0.0460	0.0591	-0.0956**
	0.0076		0.0004		-0.0132		
Western Cape ED	0.5588	0.6084	0.5623	0.5876	0.5460	0.6075	0.0120
	-0.0496		-0.0253		-0.0615		
Observations	1048	599	610	451	315	220	1061

Notes: difference in means in brackets. The final column is the estimated parameter on the non-compulsory indicator from a parametric regression discontinuity specification, only considering students that fall within 1 standard deviations of the compulsory cut-off of 50 percent in test 1.

We can similarly investigate whether or not a discontinuity in test and exam performance exists at the policy threshold. Figures A6.6 and A6.7 show similar scatter plots of average normalised test 2 and exam performance over the support of the normalised test 1 score. Students who performed just below 50 percent in the first test perform markedly higher in the second test and exam than those students who scored just above 50 percent. It is evident that there is a positive relationship between the performance in test 1 and subsequent performance throughout the semester. However, students who performed above the 50 percent in test 1 tend to perform worse in subsequent tests, excepting those who perform at the top of the distribution. The opposite is true for those students who performed below 50 percent in test 1. There therefore appears to be a degree of mean reversion in test 2 and the exam. As with tutorial attendance, different functional forms of the running variable are overlaid on the data. Inspection of the graphs prompted the use of a quadratic control function in the final model.

We now employ the parametric regression discontinuity specification from equations [6.8] and [6.9] to estimate the effect of the compulsory tutorial policy on tutorial attendance prior to test 2 and the exam. The samples under consideration are the group of students who score within one and half a standard deviation from the policy threshold. This allows for a better fit of the polynomial control function to the attendance rate over the threshold. As mentioned, linear and quadratic control functions are modelled for the first stage attendance and reduced form performance equations respectively. The results of these estimations are shown in table 6.2. Focusing first on the impact of the compulsory tutorial policy on tutorial attendance prior to test 2 and the exam, we estimate that attendance for the compulsory student group is 18 percent and 32 percent higher at the threshold prior to test 2 and the exam respectively. These estimates are statistically significant at the 1 percent level. When the window is narrowed to 0.5 of a standard deviation, the results are largely unchanged.

The inclusion of other covariates in addition to the control function does not have much of an impact on the discontinuity coefficient in the case of exam scores when observing a window of 1 standard deviation. The instrumental variable results are presented in the final column of table 6.2. A two-stage regression approach yields an estimated coefficient on tutorial attendance of 1.05, which roughly translates to a 1.5 percentage point increase in exam performance for a 10 percent increase in tutorial attendance.

Table 6.2: Regression results for tutorial attendance and performance

Within 1 standard deviation					
	First stage		Reduced form		IV (2S)
D_i	-0.3172***	-0.3179***	-0.3531**	-0.3334**	
	(0.033)	(0.032)	(0.156)	(0.144)	
Attendance					1.0503**
					(0.456)
X_i	-0.0014	-0.0005	0.0753**	0.0566**	0.0556**
	(0.002)	(0.002)	(0.030)	(0.028)	(0.027)
X_i*D_i	0.0050	0.0035	-0.0377	-0.0270	-0.0230
	(0.003)	(0.003)	(0.038)	(0.035)	(0.037)
X_i^2			0.0024	0.0018	0.0018
			(0.002)	(0.001)	(0.001)
$X_i^2*D_i$			-0.0026	-0.0020	-0.0020
			(0.002)	(0.002)	(0.002)
Other controls	No	Yes	No	Yes	Yes
Observations	947	937	947	937	937
Adjusted R^2	0.191	0.315	0.094	0.214	0.214
Within 0.5 standard deviations					
	First stage		Reduced form		IV (2S)
D_i	-0.2927***	-0.3182***	-0.5135**	-0.3265	
	(0.047)	(0.044)	(0.235)	(0.217)	
Attendance					1.0258
					(0.728)
X_i	-0.0066	-0.0027	0.1592	0.0718	0.0745
	(0.005)	(0.005)	(0.103)	(0.098)	(0.099)
X_i*D_i	0.0102	0.0100	-0.0878	-0.0365	-0.0468
	(0.008)	(0.008)	(0.119)	(0.113)	(0.110)
X_i^2			0.0113	0.0037	0.0037
			(0.010)	(0.010)	(0.010)
$X_i^2*D_i$			-0.0164	-0.0067	-0.0067
			(0.012)	(0.012)	(0.012)
Other controls	No	Yes	No	Yes	Yes
Observations	491	484	491	484	484
Adjusted R^2	0.204	0.338	0.013	0.145	0.145

Notes: *** p<0.01, ** p<0.05, * p<0.10. Robust standard errors in parentheses. Standard errors of IV estimates generated from 500 bootstraps.

As stated, local polynomial regressions are used to estimate the local treatment effect. Estimates are generated using a triangular kernel function, as well as several choice of bandwidth, namely, the Imbens and Kalyanaraman (from now on referred to as the IK bandwidth) (2009) and the McCrary (2008) ROT bandwidths. Half and twice the IK and McCrary bandwidths are used for comparison. The results are presented in table 3 below. The optimal IK bandwidth is slightly larger than the ROT bandwidth for test 2, and vice versa for exam performance. However, the results yielded by the two bandwidth choices are quite similar. The ROT bandwidth predicts a significant

increase in exam performance of 7.9 percent of a standard deviation for each additional tutorial attended, whilst the IK bandwidth yields an estimate of 10.3 percent of a standard deviation increase. Both are statistically significant at least at the 5 percent level. It is worth noting the difference in the two bandwidths upon which these estimates are based, as the narrower optimal IK bandwidth yields a larger estimated effect that is very similar to that obtained using parametric regression. This translates to a 1-1.5 percentage point increase in exam score for each additional tutorial attended.

It is further worthwhile comparing the magnitudes, statistical significance and standard errors on the estimate local treatment effect obtained under the different choices of bandwidths. It is immediately clear that the larger the bandwidth, the smaller is the estimated impact and the smaller the standard error. The contrary is true for smaller bandwidths. This is to be expected, given that a choice of larger bandwidth comes with greater precision. However, it also comes at the cost of greater bias in the estimates. Therefore, the estimates generated using the IK and ROT bandwidths may be downward biased. The following section tests the robustness of our results.

Table 6.3: Non-parametric results

		IK bandwidth	ROT bandwidth	0.5*IK bandwidth	2*IK bandwidth	0.5*ROT bandwidth	2*ROT bandwidth
	Bandwidth	*0.879*	*1.679*	*0.440*	*1.758*	*0.840*	*3.360*
1	$E[T^+] - E[T^-]$	-0.315***	-0.317***	-0.320***	-0.317***	-0.314***	-0.316***
		(0.041)	(0.030)	(0.052)	(0.029)	(0.041)	(0.026)
2	$E[Y^+] - E[Y^-]$	-0.325**	-0.250***	-0.430***	-0.246***	-0.331***	-0.227***
		(0.134)	(0.093)	(0.165)	(0.093)	(0.118)	(0.083)
2/1	LATE (Implied IV)	1.033**	0.787***	1.343**	0.775***	1.055***	0.719***
		(0.434)	(0.294)	(0.528)	(0.293)	(0.390)	(0.261)

Notes: *** p<0.01, ** p<0.05, * p<0.10. Bootstrapped standard errors generated from 500 bootstraps shown in parentheses.

6.6 Robustness Checks

One concern regarding identification of the treatment effect is that the compulsory policy may induce behavioural changes in effort. We may suspect that students who are subject to the policy are "labelled" as weak students. This may motivate compulsory tutorial students to exert more effort to better their performance relative to students who just passed and do not suffer the stigma of being a weak student. Therefore, the estimated impact of tutorial attendance may overstate the actual impact of the tutorials. One way of testing this assertion might be to analyse the behaviour of students in a subject which does not offer tutorial support. Unfortunately, such information was not readily available for this study.

Alternatively, we propose to use the second test as a potential "treatment" by comparing the average exam outcomes of "just failers" and "just passers" in the second test amongst the group of compulsory students. If we find a negative estimate, this will indicate that compulsory students who scored below 50 percent in the second test performed better in the exam than compulsory students who scored above 50 percent. Due to the fact that both groups are required to attend tutorials on a compulsory basis, and therefore receive the same "treatment", any divergence in exam performance may be ascribed to behavioural responses to "just failing" or "just passing". We used local polynomial regression to compare the average exam performance of compulsory tutorial students who scored below 50 percent in the second semester test to the average exam performance of compulsory students who scored at least 50 percent or higher. Using the optimal IK bandwidth, we estimate a LATE of -0.231. This translates to approximately an exam performance that is 2 percentage points higher for the group of compulsory students who just failed test 2, indicating that there is potentially a stronger motivation for "just failers" to pass subsequent testing relative to "just passers". However, this effect is not statistically significant.

Another issue is the potential bias in the LATE that could derive from a discontinuity in the covariates over the threshold. As mentioned, this can be tested by repeating the analysis using the residuals from a regression of the covariates (other than the control function) on performance. Alternatively, we can control for the covariates in estimation of the LATE. Both methods are employed here. We use the same optimal IK bandwidths for the regression corrected estimations as in table 6.3, and the results are shown in columns 2 and 3 of table 4 below. Correction for covariates reduces the estimated effect of tutorial attendance. This may suggest violation of the continuity assumption of one or more of the covariates. A visual inspection of local linear regression graphs for each covariate indicates no significant discontinuity in the covariates at the threshold (see figures A6.8-A6.26 of the appendix), except in the case of "white race group" where we find a significantly higher (at the 10 percent level) density of non-compulsory students than compulsory students close to the cut off. However, this discontinuity disappears with a smaller bandwidth. The reduced LATE could also suggest a discontinuity in one or more of the unobservables that may be related to the observable characteristics, such as ability and effort. Comparisons of the estimates from table 3 with the regression corrected estimates in table 4 indicate that the results are not statistically significant; therefore we can conclude that inclusion of the covariates in the non-parametric model results in consistent estimates of the LATE and has improved precision as evidenced by smaller standard errors.

Students were permitted to leave the compulsory tutorial programme if they were able to score at least 65 percent in the second semester test. As a result, 56 of the 599 compulsory students

were no longer compelled to attend the tutorials. The results may be biased to the inclusion of this group of students as their behaviour may have been altered before test 2 (more motivated to leave the programme) and before the exam (refrained from attending tutorials on a regular basis). However, students were only made aware of their performance in test 2 in the 9th week of tutorials, therefore only leaving 3 of the 5 remaining compulsory tutorials optional for this group of students. As a result, only 9 of these 56 students did not attend at least 4 of the 5 compulsory tutorials between test 2 and the exam. The LATE on the exam was re-estimated for two sub-samples of students: sample excluding all compulsory students who scored at least 65 percent in test 2; and a sample excluding only those compulsory students who scored at least 65 percent in test 2 and "left" the programme. The results of these estimations based on the same IK bandwidth from table 3 are shown in columns 4 and 5 of table 6.4.

Excluding all students who achieved at least 65 percent in test 2 dramatically reduces the local treatment effect to 0.46. The effect is also non-significant. Excluding only those students who "left" the programme reduces the estimate slightly. This result may be of concern, as it suggests that the tutorials only had impact (so to speak) for a relatively small group of students; that is, those students who failed test 1, but performed well in test 2. The question then becomes: is this group of students different to the other compulsory students? Comparison of average observables indicates that this group of students tend to have a significantly higher proportion of students enrolled in accounting and actuarial science, as well as higher average performance in matric mathematics. This suggests that this group of students are most likely more able than the other compulsory students, and may have been more motivated to pass in future tests. The positive impact of tutorial attendance may therefore mask a change in behaviour that is policy driven. However, without other information with which we could test how student effort changes in response to this policy, it is difficult to say how much of the positive effect is due to motivational factors and that which is due to the tutorials. On the other hand, the fact that only 9 of the 56 students decided to exit the compulsory tutorial programme suggests that even these higher performing students attached value to being exposed to the compulsory tutorial programme.

Finally, the result may also be sensitive to the choice of exam score used. The dependent variable includes exam scores from both the first and second exam papers. Compulsory students were no more likely to opt to write the second exam option than non-compulsory students. However, we may be concerned that the two papers were of different quality. Furthermore, students who wrote the second exam may have had access to the first exam paper, which may have benefited them. We therefore re-estimate the LATE excluding those students who only wrote the

second exam option. The results are shown in the final column of table 4. Exclusion of this group of students has no significant effect on the predicted LATE.

Table 6.4: Sensitivity checks

		(1) Compulsory students: test 2 as treatment	(2) Regression corrected (residual)	(3) Regression corrected (inclusion of covariates)	(4) Excluding compulsory students who scored >=65% in test 2	(5) Excluding compulsory students who scored >=65% in test 2 & left programme	(6) Excluding students who only wrote exam 2
1	$E[T^+] - E[T^-]$	-	-0.304***	-0.318***	-0.336***	-0.331***	-0.305***
			(0.038)	(0.034)	(0.040)	(0.038)	(0.040)
2	$E[Y^+] - E[Y^-]$	-0.2312	-0.266**	-0.301**	-0.154	-0.270**	-0.289**
		(0.211)	(0.121)	(0.116)	(0.126)	(0.119)	(0.135)
2/1	LATE (Implied IV)	-	0.874**	0.946**	0.456	0.817**	0.946**
			(0.416)	(0.372)	(0.372)	(0.352)	(0.456)

Notes: *** p<0.01, ** p<0.05, * p<0.10. Bootstrapped standard errors generated from 500 bootstraps shown in parentheses.

6.7 Conclusion

The poor academic performance and retention of undergraduate students has prompted the adoption of alternative methods of learning and teaching that not only provide the necessary support to students and enhance their learning approaches, but are also cost-effective. The literature has provided mixed results regarding the impact of peer tutoring on the academic performance of undergraduate students (Topping, 1996). Although much of the existing research includes a cross-sectional component that typically compares the performance of students who have had tutoring versus those who have not, efforts have been made towards the adoption of quasi-experimental and random control designs that include both cross-sectional and longitudinal components that can control for potentially confounding factors or eliminate the sample specific biases that explain the observed effects.

This study aimed to contribute to the literature through the use of a fuzzy regression discontinuity design that potentially corrects for the issue of selection on unobservables that may bias the point estimates of tutorial attendance. The local average treatment effect is estimated using a bandwidth of observations around the policy threshold of 50 percent in the first semester test. It is clear that the policy significantly increases the tutorial attendance amongst compulsory tutorial students following the first semester test. IV regression results indicate a positive impact of tutorial attendance on test 2 performance. A 10 percent increase in tutorial attendance results in

approximately a 10 percent standard deviation increase in exam performance. However, this result is only statistically significant in the case of the latter. Quantitatively similar impacts are found using local linear polynomial regression, although the results are sensitive to choice of bandwidth and specification of the control function. Robustness checks indicate that the results are fairly insensitive to the inclusion of the other covariates. However, the exclusion of the best performing compulsory students who were permitted to leave the programme decreases the treatment effect. This raises the concern that the result may be biased by unobservable factors such as motivation and effort that are not exogenous to the tutorial policy. Nevertheless, the fact that only 9 of the 56 students took advantage of the exit option indicates that the students themselves attach value to attending these tutorials.

In conclusion, being assigned to the compulsory tutorial programme does affect performance but only for students that seem to have the ability to perform anyway. Unfortunately, this study is not able to unpack the mechanism through which assignment to the compulsory tutorial programme impacts on these students. The analysis would have benefited greatly through the inclusion of additional information, unavailable to the authors at the time of this study, regarding the performance of students in other coursework besides microeconomics where such interventions are not currently in place, as well as attitudinal and behavioural changes towards class attendance and time spent in studying. The longitudinal aspect of these types of programmes also needs to be considered, as the benefits of peer tutoring may only emerge at a later stage, or may even be short-lived. Differences between tutored and untutored students may either decline or increase over time depending on the adaption strategies of individual students (Jacobi, 1991).

Appendix to Chapter 6

Figure A6.1: Student attendance prior to test 1

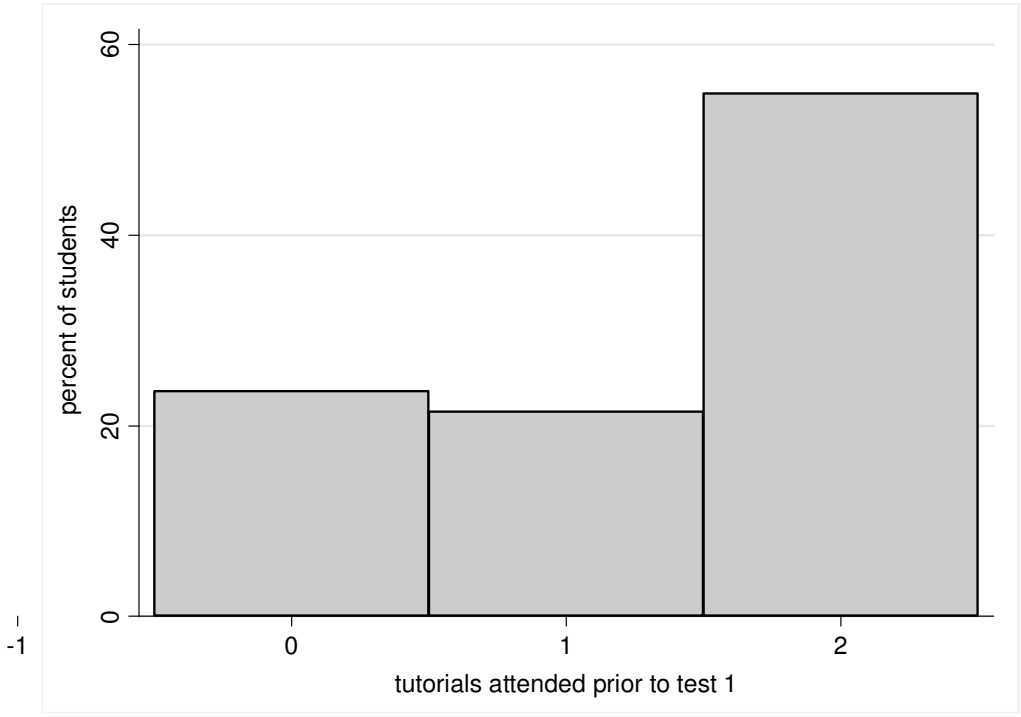


Figure A6.2: Student attendance by treatment

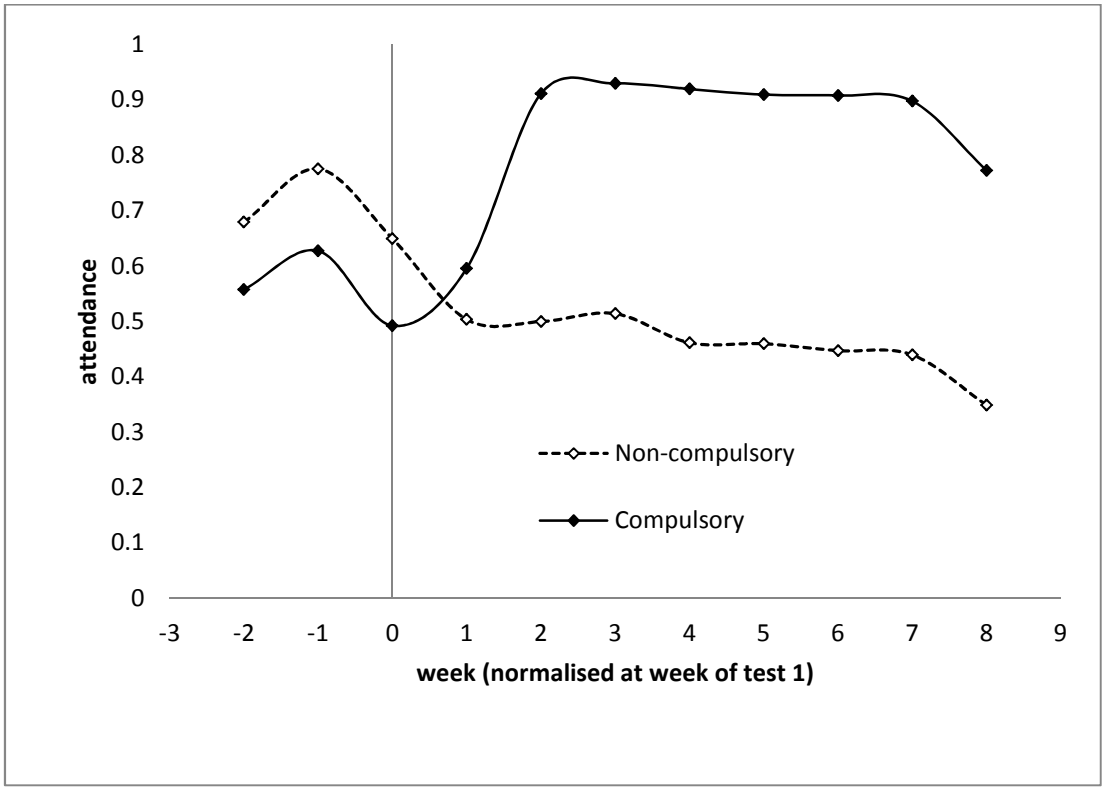


Figure A6.3: Student attendance prior to test 1

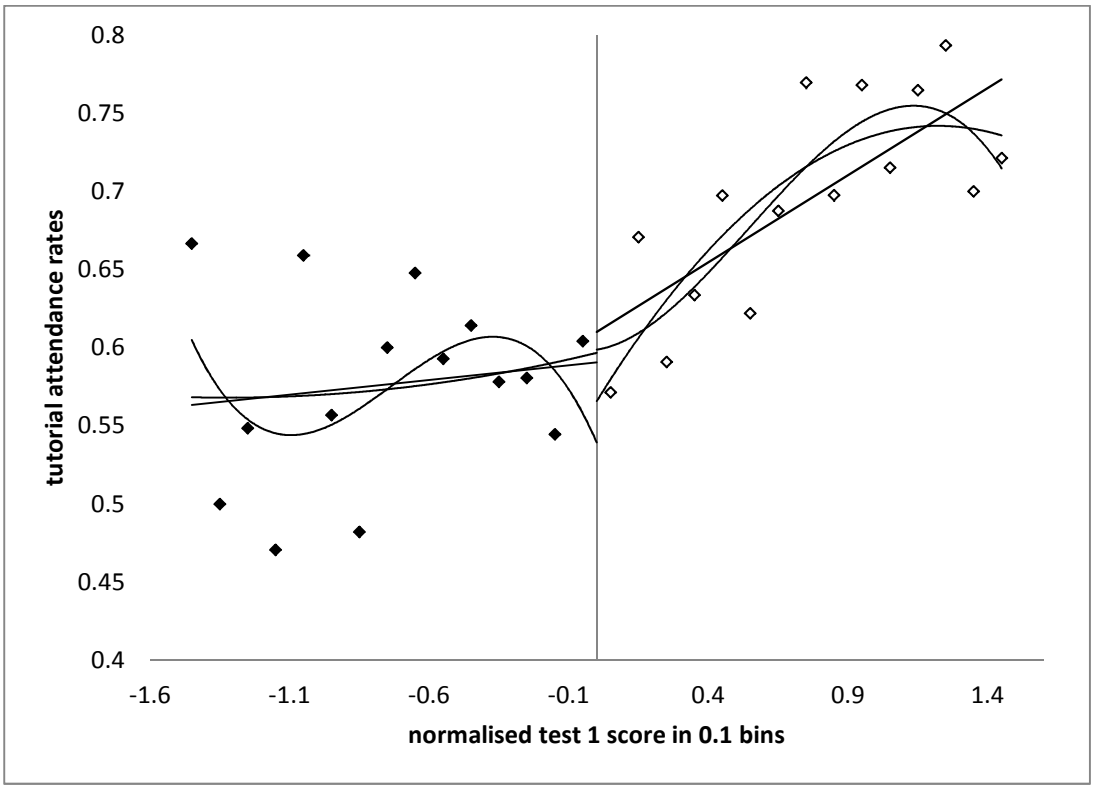


Figure A6.4: Student attendance prior to test 2

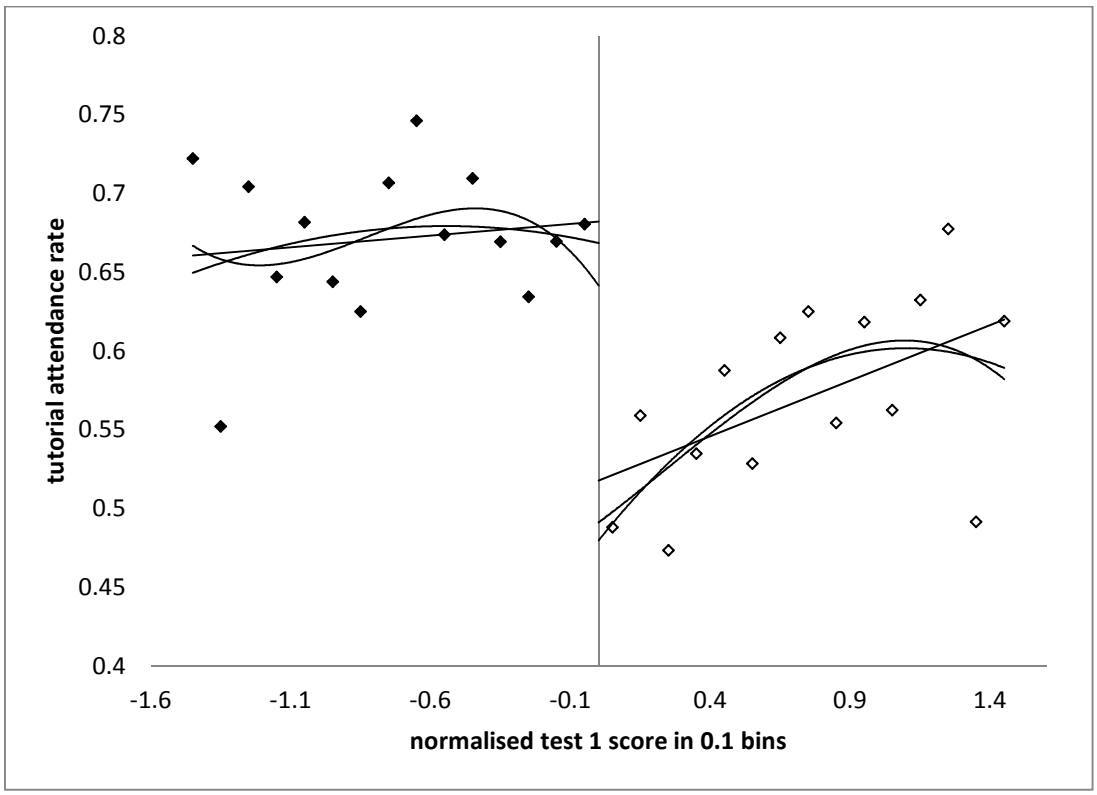


Figure A6.5: Student attendance prior to exam

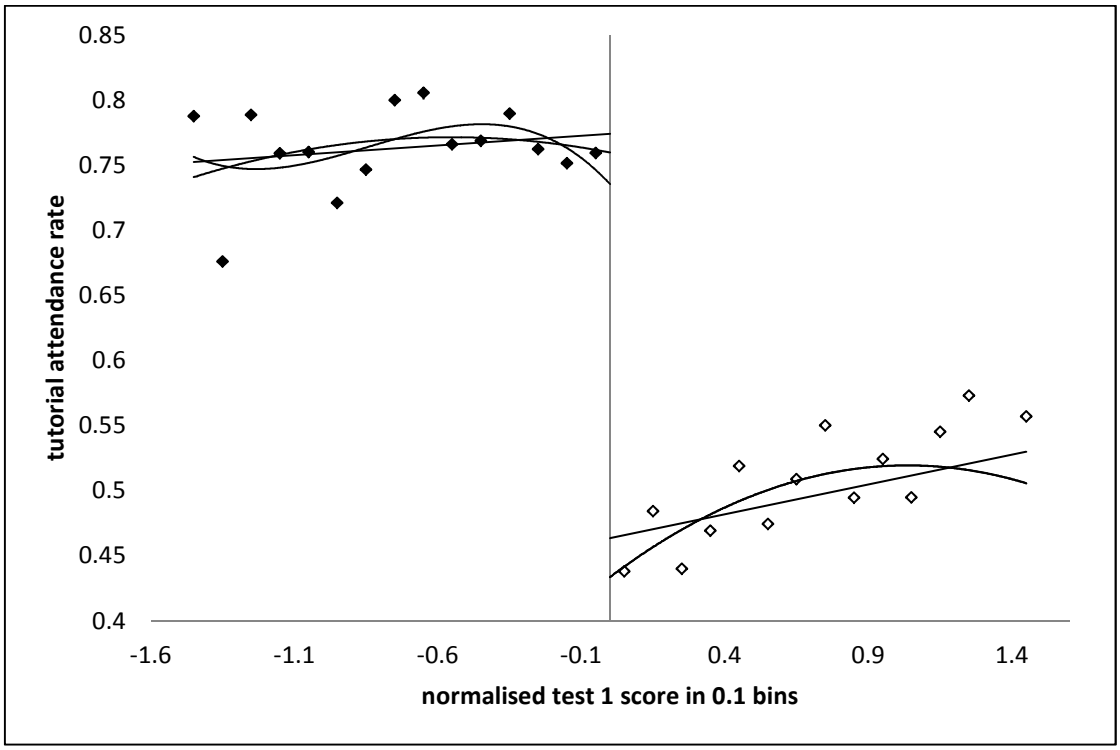


Figure A6.6: Student performance in test 2

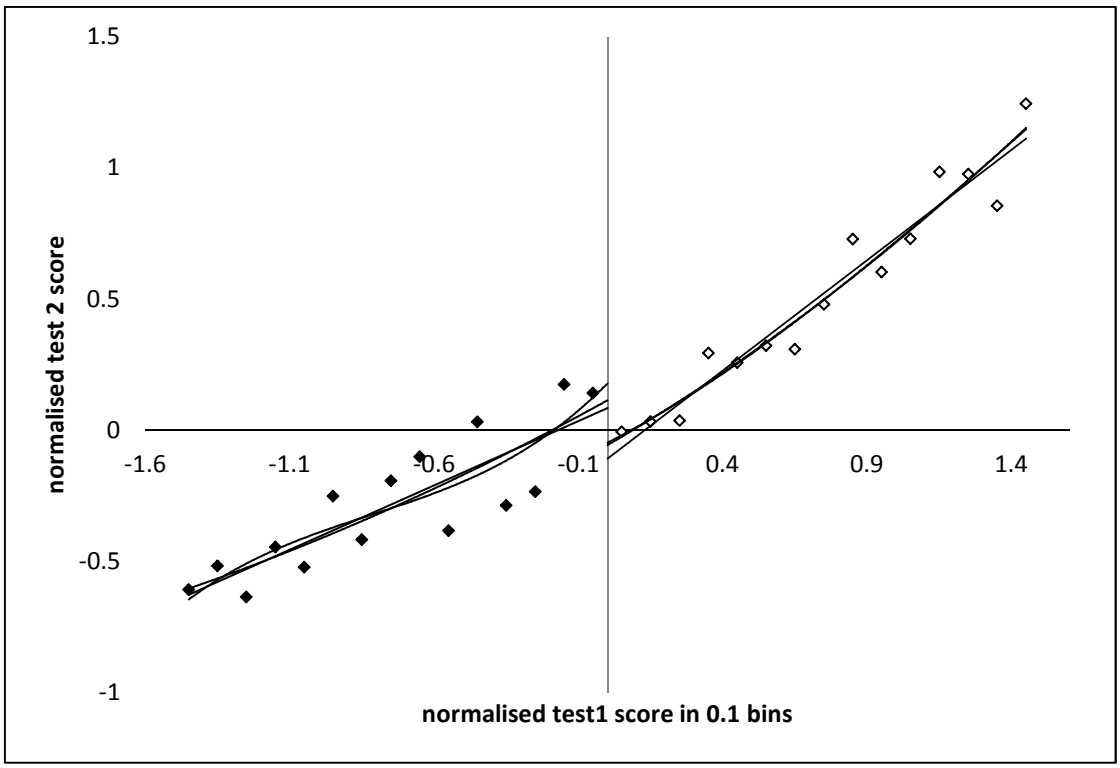


Figure A6.7: Student performance in exam

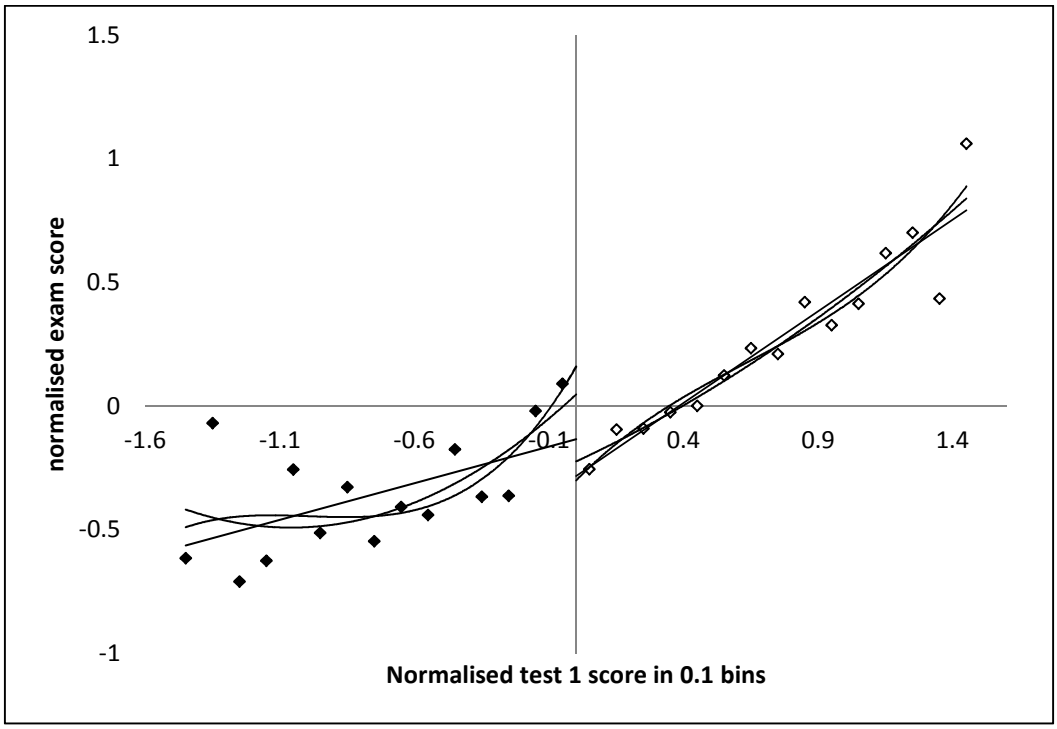


Figure A6.8

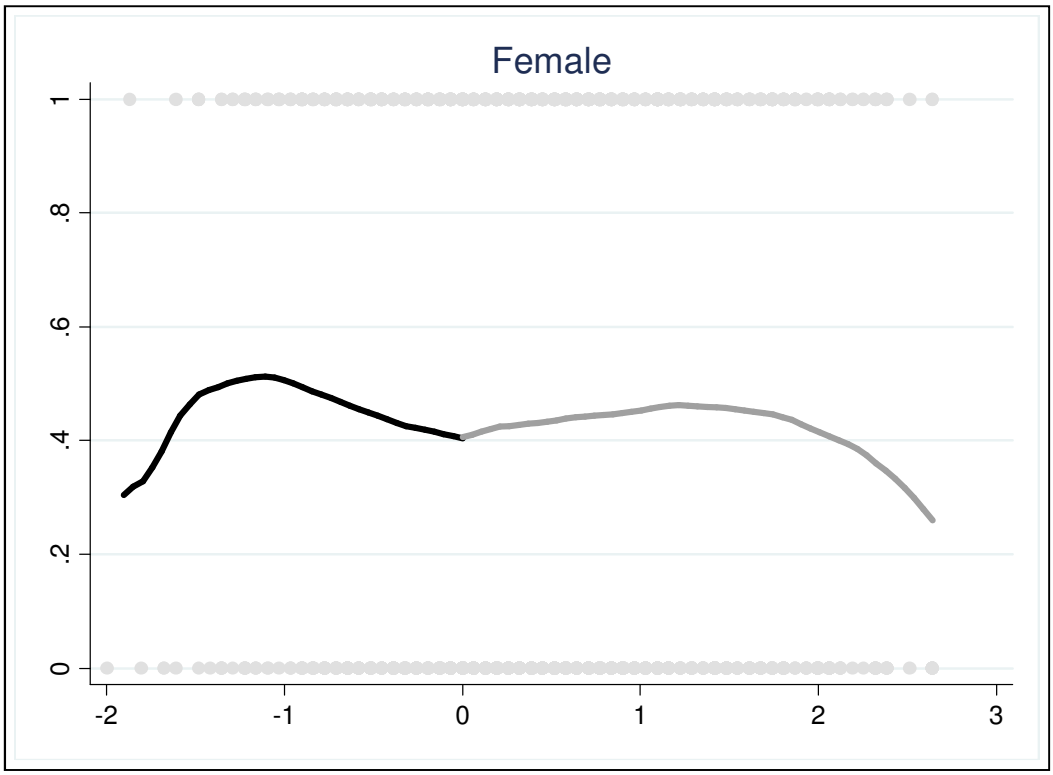


Figure A6.9

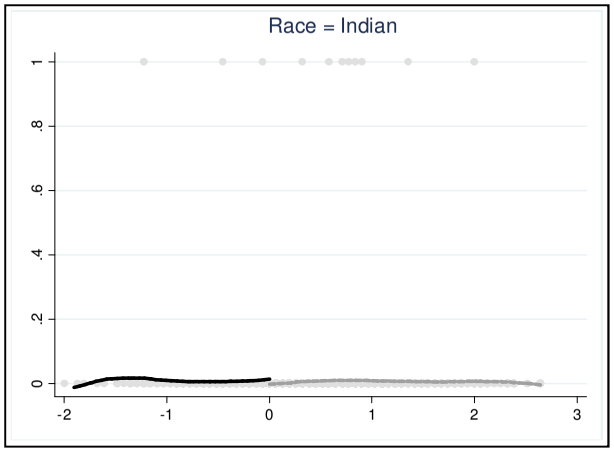


Figure A6.10

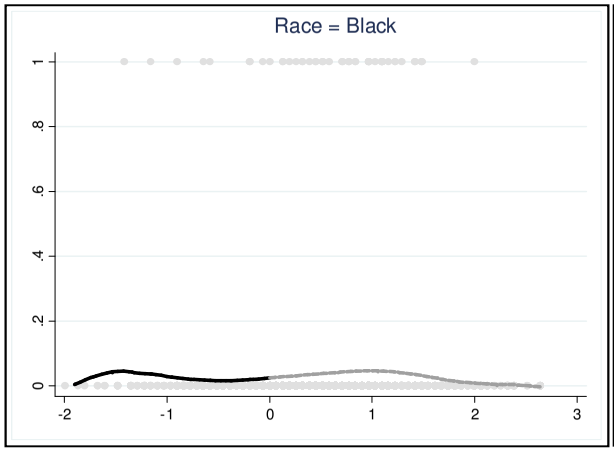


Figure A6.11

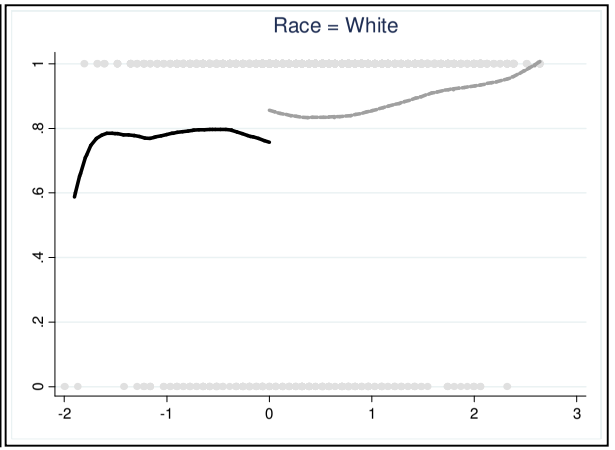


Figure A6.12

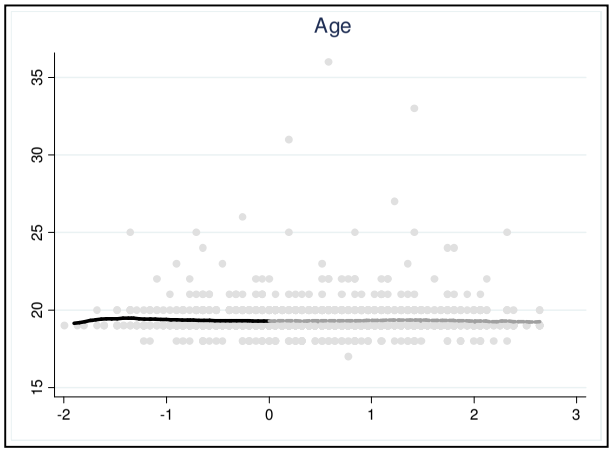


Figure A6.13

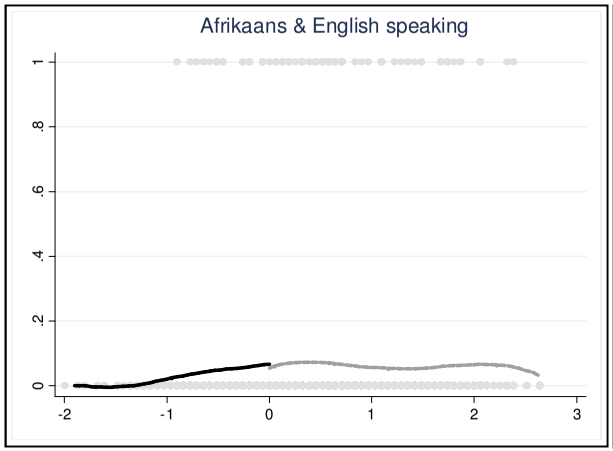


Figure A6.14

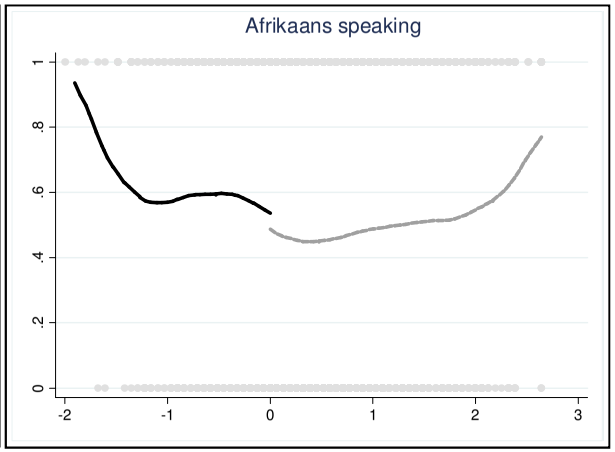


Figure A6.15

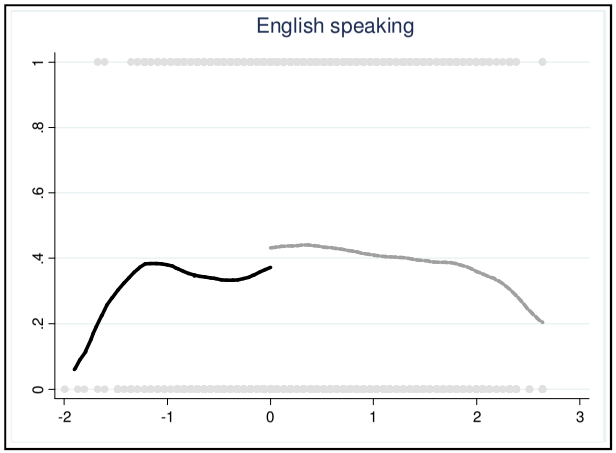


Figure A6.16

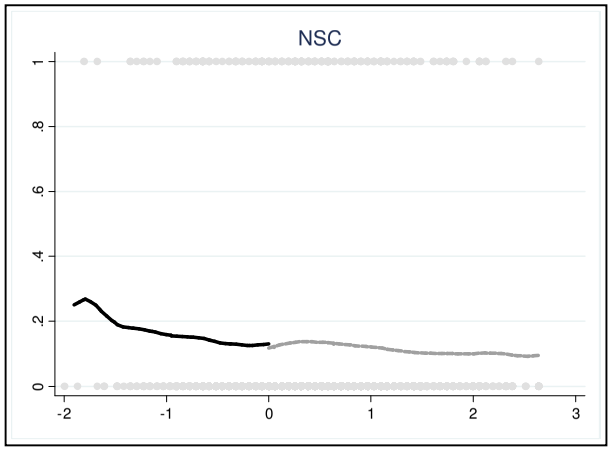


Figure A6.17

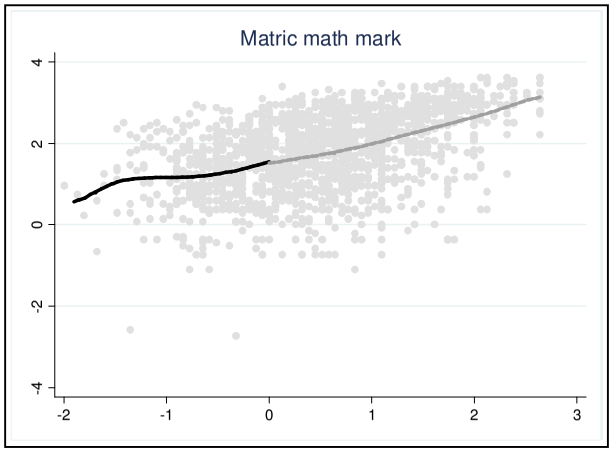


Figure A6.18

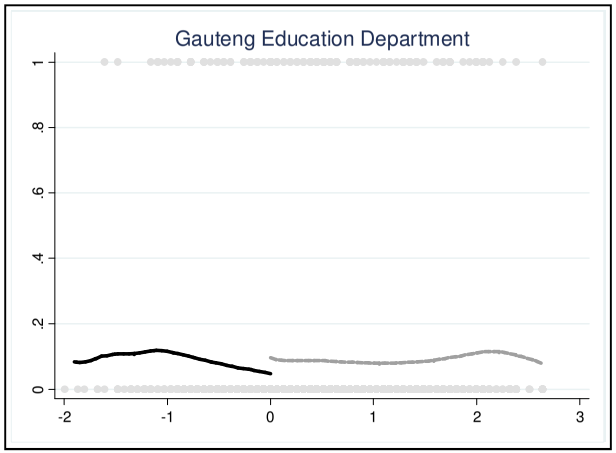


Figure A6.19

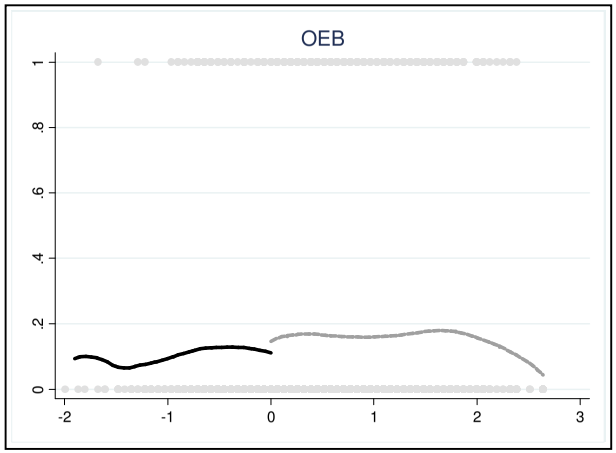


Figure A6.20

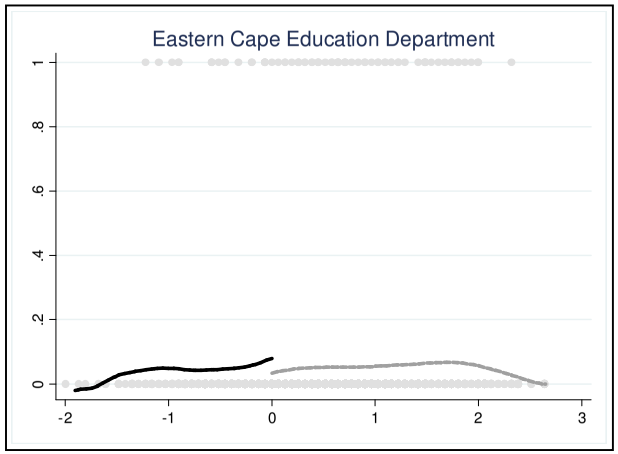


Figure A6.1: Alternative specifications of the control function

	First stage: tutorial attendance				Reduced form: Test 2 performance			
	(1 SD)		(0.5 SD)		(1 SD)		(0.5 SD)	
Linear control function								
D_i	-0.3172***	-0.3179***	-0.2732***	-0.3123***	-0.2095**	-0.2202**	-0.2769*	-0.2335*
	(0.033)	(0.032)	(0.048)	(0.045)	(0.106)	(0.099)	(0.148)	(0.139)
X_i	-0.0014	-0.0005	-0.0078	-0.0027	0.0321***	0.0247***	0.0506**	0.0362*
	(0.002)	(0.002)	(0.005)	(0.005)	(0.008)	(0.007)	(0.020)	(0.020)
$D_i.X_i$	0.0050	0.0035	0.0121	0.0107	0.0019	-0.0003	-0.0233	-0.0285
	(0.003)	(0.003)	(0.009)	(0.008)	(0.010)	(0.009)	(0.027)	(0.026)
R^2	0.193	0.330	0.209	0.367	0.097	0.233	0.019	0.185
Quadratic control function								
D_i	-0.2923***	-0.3050***	-0.2771***	-0.2922***	-0.3531**	-0.3334**	-0.5135**	-0.3265
	(0.049)	(0.046)	(0.069)	(0.066)	(0.156)	(0.144)	(0.235)	(0.217)
X_i	-0.0099	-0.0086	-0.0031	-0.0128	0.0753**	0.0566**	0.1592	0.0718
	(0.007)	(0.007)	(0.023)	(0.024)	(0.030)	(0.028)	(0.103)	(0.098)
X_i^2	-0.0005	-0.0004	0.0005	-0.0010	0.0024	0.0018	0.0113	0.0037
	(0.000)	(0.000)	(0.002)	(0.002)	(0.002)	(0.001)	(0.010)	(0.010)
$D_i.X_i$	0.0138	0.0157	0.0041	0.0184	-0.0377	-0.0270	-0.0878	-0.0365
	(0.012)	(0.011)	(0.034)	(0.033)	(0.038)	(0.035)	(0.119)	(0.113)
$D_i.X_i^2$	0.0004	0.0002	-0.0001	0.0013	-0.0026	-0.0020	-0.0164	-0.0069
	(0.001)	(0.001)	(0.004)	(0.003)	(0.002)	(0.002)	(0.012)	(0.012)
R^2	0.194	0.331	0.209	0.367	0.099	0.234	0.022	0.186

Table A6.1 continued: Alternative specifications of the control function

	(1 SD)		(0.5 SD)		(1 SD)		(0.5 SD)	
	Cubic control function							
D_i	-0.2821***	-0.3370***	-0.3390***	-0.4397***	-0.4069*	-0.2777	-0.1529	-0.1855
	(0.064)	(0.059)	(0.110)	(0.106)	(0.211)	(0.195)	(0.422)	(0.398)
X_i	-0.0132	0.0096	0.0395	0.1175	0.1189	0.0613	-0.1970	-0.0947
	(0.020)	(0.020)	(0.085)	(0.085)	(0.081)	(0.077)	(0.351)	(0.340)
X_i^2	-0.0009	0.0019	0.0106	0.0300	0.0081	0.0024	-0.0744	-0.0363
	(0.002)	(0.002)	(0.020)	(0.020)	(0.010)	(0.009)	(0.080)	(0.079)
X_i^3	-0.0000	0.0001	0.0007	0.0021	0.0002	0.0000	-0.0058	-0.0027
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.005)	(0.005)
$D_i.X_i$	0.0136	-0.0014	0.0013	-0.0828	-0.0965	-0.0817	0.2770	0.2002
	(0.029)	(0.028)	(0.102)	(0.100)	(0.096)	(0.090)	(0.375)	(0.360)
$D_i.X_i^2$	0.0014	-0.0023	-0.0224	-0.0387	-0.0061	0.0043	0.0667	0.0116
	(0.004)	(0.003)	(0.025)	(0.024)	(0.012)	(0.011)	(0.088)	(0.087)
$D_i.\ X_i^3$	-0.0000	-0.0001	0.0002	-0.0014	-0.0003	-0.0003	0.0060	0.0043
	(0.000)	(0.000)	(0.002)	(0.002)	(0.000)	(0.000)	(0.006)	(0.006)
R^2	0.194	0.331	0.210	0.368	0.100	0.235	0.096	0.187
Other controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	947	937	491	484	947	937	491	484

Notes: *** p<0.01, **p>0.05, *p<0.10. Robust standard errors generated from 500 bootstraps shown in parentheses.

Chapter 7

Summary of main findings

The enduring gap in the quality of education, attainment and performance that persists between the former (white) advantaged, well-functioning, mainly affluent schooling system and the majority disadvantaged, mainly black dysfunctional school system can be viewed as having its origins in the highly unequal forms of provision and expenditure that existed during the apartheid regime. Although policy reforms and legislation since 1994 have contributed to equal distributions of expenditures across provinces and the provision of zero-fee compulsory schooling up to grade 9, there is evidence to suggest that this spending could be more pro-poor. Furthermore, the private funding available to the wealthiest schools in the form of school fees and fund raising results in dramatically different quality of schooling inputs and processes being provided to students attending these schools. The result of this has been an influx of chiefly middle-class black children (who can afford the high school fees) into the former advantaged school system. The main research question of this thesis can therefore be summarised as: "what are the main contributing factors to differences in school quality and effectiveness across the South African schooling system?"

The introductory chapter of this thesis introduced a social justice perspective to researching education quality which assigns a central role to the context under which teaching and learning takes place. Tikly (2011:11) makes the argument that a deeper appreciation of context is required in order to characterise quality education, as it "encourages policy makers to take cognisance of changing national development needs, the kinds of schools that different students attend and the forms of educational disadvantage faced by different groups of learners when considering policy options". Consideration of educational quality through a social justice framework introduces a methodological challenge. Specifically, research needs to recognise the complex and multi-dimensional nature of the issues relating to the quality of education and how they impact on different groups of students, particularly those from disadvantaged home backgrounds. The analysis and research questions posed by chapters 2 to 6 of this thesis aimed to go beyond the standard quantitative techniques and introduce inter-disciplinary and relatively under-utilised methodological approaches in education to assess schooling effectiveness in South Africa. These techniques were chosen with the intention of being both sensitive to context and internationally relevant.

Modelling school effectiveness within former disadvantaged South African public schools

Chapter 2 adopted machine learning techniques for modelling school effectiveness (production process) within former black African and homeland primary schools. The methodological approach was chosen specifically because it allows for the more effective modelling of complex relationships, such as is observed in education. The NSES 2008 grade 4 dataset containing data on student, household and school level characteristics as well as identifies former school department was employed. The robustness of the empirical approach was tested against multiple data sets and using alternative parametric and non-parametric approaches.

The results indicate that social contexts are relevant for determining student outcomes. As reflected by figure 1.1, the right blend (interaction) of enabling processes and schooling inputs at the levels of national policy, school and the home/community is vital for achieving the desired schooling outcomes. The findings of the regression tree analysis indicate that classroom processes, particularly time-on-task and opportunity-to-learn, play dominant roles in determining performance, particularly through the way they interact with other teacher characteristics and home background factors. Results from a mixed effects random forest model further stresses this. Less affluent South African schools face constraints that inhibit effectiveness. The socio-economic context of students and the communities from which they come are particularly binding as they not only limit the opportunities for supplementary learning outside of school, but also inhibit learning at school (through, for example, poor nutrition) as well as limit the role of parental “voice” which can contribute to school accountability.

The findings of this chapter therefore suggest that the most significant positive interventions for the black school system would be of the type that affect enabling inputs and processes, and work to overcome the gaps that often exist between schools, households/communities, and national policy. This includes the professional development of teaching staff and school principals to understand, choose, develop and evaluate relevant and effective practices within the context of their own school’s status and culture. In spanning the learning gap that exists between the school and home environments, a better understanding of those classroom processes that disproportionately advantage poor students is required. This may include extending the amount of in-school learning time for children who lack the necessary supporting inputs at home.

Estimating the impact of attending a former advantaged school

The focus of chapters 3 and 4 was to estimate the treatment effect (impact) on student performance children as a result of attending a former advantaged (white) school. In assessing this effect, it needs

to be understood that the "average" or typical South African student does not exist in any conceivable way that can permit the treatment effect to be identified simply through a comparison of average performance across the two schooling sub-systems (that is, former disadvantaged and former advantaged). Selection into school type is driven by predominantly household background factors that allow for (i) ease of mobility to locate near to better schools and (ii) financial capacity to afford the higher school fees. These are also likely to be highly correlated to race and region. Identification of black children across the two school systems might serve as a good comparator group, although this characteristic is almost never available within the observational data. Furthermore, there is limited homogeneity amongst black learners across the two schooling systems.

In order to control for the selection bias intrinsic to school choice in South Africa, the analysis of chapters 3 and 4 made use of the PIRLS 2006 and prePIRLS 2011 datasets that capture a wealth of student and home background characteristics. Given that student testing within these datasets was furthermore conducted using all 11 official languages, the former department of the school (also rarely identified in observational data) could be proxied by the language of learning and teaching at the school. The methodological approaches adopted by the analyses of chapter 3 and 4 incorporate selection on observables in different but related ways.

Chapter 3 uses a semi-parametric form of the well-known Oaxaca-Blinder decomposition to estimate the effect of attending an English/Afrikaans testing school. Specifically, propensity score reweighting was used to construct suitable counterfactuals so that the average reading test score gap between English/Afrikaans testing and African language testing schools could be decomposed into three components: (1) the explained gap that is driven by differences in student and home background characteristics; (2) the school resource gap that is driven by differences in the distribution of school resources (including school SES); and (3) the school efficiency gap that is due to differences in school effectiveness (processes). The explained gap was estimated to account for roughly 40-60% of the total performance gap, whilst the school resource and school efficiency gaps were estimated to account for 14-36% and 14-26%, respectively. Whilst home background plays a dominant role in determining outcomes across school systems, successfully addressing inequalities in the distribution of school inputs and processes that augment performance as well as inequalities in school quality (effectiveness) may as much as halve the average performance gap between the two former school departments.

Chapter 4 similarly makes use of propensity score reweighting to estimate the treatment effect of attending a former advantaged school, although emphasis is placed on the "marginal" student; that is, the South African student who is potentially closer to the margin of attending a

former advantaged versus a former disadvantaged school. Estimation therefore focused on the *local* average treatment effect of attending an English/Afrikaans testing school (as proxy for the former advantaged school department). Matching and balancing weights are used to fully account for selection on pre-treatment covariates. The findings indicate that, as with chapter 3, home background accounts for roughly 40 percent of the average test score gap between grade 4 students attending English/Afrikaans testing and African language testing schools. The local average treatment effect of school type is estimated to be approximately equivalent to 1 to 1.3 years of learning, or 0.5-0.7 standard deviations. This estimate of school type is of the same magnitude as Coetzee (2014) who uses the grade 4-5 National School Effectiveness Survey panel data to estimate a value-added model of attending a former white school.

The findings of chapters 3 and 4 therefore show that whilst the circumstances of a child's home background plays a significant role in determining school performance, we cannot ignore the fact that the quality of school attended plays an equally important role in explaining the bimodal distribution of performance in the South African school system. Policy targeted specifically at improving the quality of schools, whilst taking cognisance of the social context of schools and their students can therefore do much to improve educational outcomes and, more generally, the enhancement of human capabilities.

The final contribution of chapter 4 relates to the methodology implemented which illustrated that regression analysis can be utilised for estimating the school type effect if the conditional independence and common support assumptions are satisfied i.e. a fully saturated regression model is used. Even if a researcher opts to use a non-parametric weighting or matching technique instead, this should be combined with regression in the manner of a doubly robust estimator.

The effect of teacher knowledge on learning outcomes

As was revealed by the analysis of chapter 2, teachers play a central role in learning. The impact of teacher quality in South Africa is not well understood, at least on a nationally representative level. The majority of studies that have placed explicit focus on teacher knowledge and student performance either have limited external validity (as they are limited to small scale regional studies) or are focused purely on mathematics.

This study adds to the debate of the determinants of student performance in South Africa through identifying the impact of teacher subject knowledge as well as other teacher (e.g. education and experience) and classroom (e.g. textbook availability) factors on grade 6 student performance in

reading and mathematics. The rich 2007 SACMEQ dataset was employed with correlated random effects model estimation in order to estimate the causal link between teacher test scores and student test scores. The results indicate that overlooking the selection and omitted variable (endogeneity) biases that exist when modelling schooling data can lead to upwardly biased estimates of the effect of teacher knowledge on performance.

Although no significant effect of teacher subject knowledge was estimated for the full sample, separation by school wealth quintile (as a proxy for former department) indicated heterogeneous effects across the school system. Significant positive and non-linear effects of teacher subject knowledge were estimated for the wealthiest quintile of schools, whilst no significant effect of teacher knowledge was estimated for the poorest four school wealth quintiles. A similar result was found for teacher education. However, the large and highly insignificant effect of young and inexperienced teachers in poor schools may signal the better quality of training received by teachers who have most recently entered the teaching profession. Other policy relevant findings from chapter 5 include a large and significant effect size of textbook provision in poor schools which outweighs the effect sizes of all other observable teacher and classroom characteristics.

These large positive and significant effects of teacher education and experience are dissimilar to those typically found in the South African literature and could be related to teacher unobservable quality. Once teacher unobervables were corrected for through the use of a sample-teacher both-subject sample, a positive effect size of teacher knowledge on performance of approximately 13-15 percent of a standard deviation and 5-6 percent of a standard deviation was estimated for the poorer subset and wealthier subset of South African schools, respectively. These estimates are in line with international findings that adopt similar techniques for estimating teacher effects. One of the main conclusions of this chapter was that factors contributing to effective teaching such as high quality training, pedagogical skill and opportunity to teach appear to be lacking amongst the poorer part of the South African education system,. The results also suggest that teachers within these schools may be working under conditions that hinder the transmission of knowledge to students, such as mismanagement, poor instructional leadership and poor teacher collaboration. The finding that the estimated effect size of teacher knowledge is of twice the magnitude in the poorest subset of schools reflects the relative importance of teacher knowledge for learning across the school system.

The impact of peer tutoring as a higher education learning intervention

The poor academic performance of students in South Africa is not only reserved for basic education. The high rate of dropout amongst undergraduate students in universities has urged academic department to adopt alternative methods of learning and teaching that learning and are cost-effective. Efforts have been made within the literature towards the adoption of quasi-experimental and random control designs that can control for potentially confounding factors or eliminate sample specific biases. The analysis conducted in this chapter aimed to contribute to the literature through the use of a fuzzy regression discontinuity design that potentially corrects for the issue of selection on unobservables that may bias the point estimates of tutorial attendance. The compulsory tutorial programme delivered by the Economics Department of Stellenbosch University was analysed.

The local average treatment effect was estimated using a bandwidth of observations around the policy threshold of 50 percent in the first semester test. Instrumental variable regression results indicated a positive but insignificant impact of tutorial attendance on test 2 performance. A statistically significant 10 percent increase in tutorial attendance was found to lead to roughly a 10 percent standard deviation increase in exam performance. Quantitatively similar impacts were found using local linear polynomial regression. Robustness checks indicated that whilst the results were fairly insensitive to the inclusion of the other controls, the exclusion of the best performing student who were able to “opt out of” the programme after the second test resulted in a smaller and statistically insignificant treatment effect. Therefore, the compulsory tutorial programme appears to have some effect on performance but only for those students that seem to have the ability to perform anyway. Future research would need to unpack the mechanism through which assignment to the compulsory tutorial programme impacts on these students.

Bibliography

Abadie, A., and Imbens, G. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1).

Altinok, N. (2013). The impact of teacher knowledge on student achievement in 14 Sub-Saharan African countries. *Background paper for EFA Global Monitoring Report*, *4*.

Ammermüller, A. (2006). PISA: What makes the difference*? Empirical Economics*, 33(2), 263–287.

Ammermüller, A., and Dolton, P. J. (2006). Pupil-teacher gender interaction effects on scholastic outcomes in England and the USA (No. 06-06). ZEW Discussion Papers. ZEW - Center for European Economic Research.

Angrist, J. D., and Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ashenfelter, O., and Zimmerman, D. J. (1997). Estimates of the returns to schooling from sibling data: Fathers, sons, and brothers. *Review of Economics and Statistics*, 79(1), 1–9.

Austin, P. C., Lee, D. S., Steyerberg, E. W., and Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? *Biometrical Journal*, 54(5), 657–73.

Baker, D. (1998). *The Implementation of Alternative Assessment Procedures and Washington State Educational Reform*.

Bang, H., and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–73.

Barrera-Osorio, F., Garcia-Moreno, V., Patrinos, H.A. and Porta, E. (2011). Using the Oaxaca-Blinder decomposition technique to analyze learning outcomes changes over time: an application to Indonesia's results in PISA mathematics (No. 5584). World Bank Policy Working Papers. World Bank.

Barrett, A. M. (2007). Beyond the polarization of pedagogy: models of classroom practice in Tanzanian primary schools. *Comparative Education*, 43(2), 273–294.

Barry, S., and Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43(3), 413–423.

Barsky, R. B., Bound, J., K.K., C., and J.P., L. (2002). Accounting for the black-white wealth gap: A nonparametric approach. *Journal of the American Statistical Association*, 97, 663–673.

Battistich, V., Solomon, D., and Kim, D. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American Educational Research Association*, 32(3), 627–658.

Bauer, E., and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36, 105–139.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *The journal of political economy*, 9-49.

Bedi, A. S., and Marshall, J. H. (2002). Primary school attendance in Honduras. *Journal of Development Economics*, 69, 129–153.

Behrman, J. R., Ross, D., and Sabot, R. (2008). Improving quality versus increasing the quantity of schooling: Estimates of rates of return from rural Pakistan. *Journal of Development Economics*, 85, 94–104.

Berk, R. A. (2008*). Statistical learning from a regression perspective*. Springer Science & Business Media.

Bickel, P. J., and Kwon, J. (2001). Inference for semiparametric models: some questions and an answer. *Statistica Sinica*, 863-886.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8, 436–455.

Botezat, A., and Seiberlich, R. R. (2013). Educational performance gaps in Eastern Europe. *Economics of Transition*, 21(4), 731–756.

Boud, D., Cohen, R., and Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education*, 24, 413–426.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*, The Wadsworth Statistics and Probability Series, Wadsworth International Group, Belmont California (pp. 356).

Burger, R. (2011). School effectiveness in Zambia: The origins of differences between rural and urban outcomes. *Development Southern Africa*, 28(2), 157–176.

Bush, T., Joubert, R., Kiggundu, E., and van Rooyen, J. (2010). Managing teaching and learning in South African schools. *International Journal of Educational Development*, 30, 162–168.

Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5), 885-897.

Carnoy, M., and Arends, F. (2012). Explaining mathematics achievement gains in Botswana and South Africa. *PROSPECTS*, 42(4), 453–468.

Carnoy, M. and Chisholm, L. (2008). Towards Understanding Student Academic Performance in South Africa: A Pilot Study of Grade 6 Mathematics Lessons in Gauteng Province. Retrieved from http://www.hsrc.ac.za/en/research-outputs/view/3743

Cattaneo, M. A., and Wolter, S. C. (2012). Migration policy can boost PISA results: Findings from a natural experiment (No. 6300). IZA Discussion Papers. Institute for the Study of Labor (IZA).

Chapman, J. W., and Tunmer, W. E. (2003). Reading difficulties, reading-related self-perceptions, and strategies for overcoming negative self-belief. *Reading & Writing Quarterly*, 19(1), 5–24.

Chetty, M. and Moloi, M. Q. (2011). The SACMEQ III project in South Africa: A study of the conditions of schooling and the quality of education.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2012). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 6, 266–298.

Chisholm, L. (2004). *Changing class: Education and social change in post-apartheid South Africa* (p. 340). Zed Books Ltd.

Chisholm, L. (2009). *An overview of research, policy and practice in teacher supply and demand*, 1994–2008 (p. 56). HSRC Press.

Chisholm, L., Motala, S., and Vally, S. (2003). *South African education policy review*, 1993-2000 (p. 850). Johannesburg: Heinemann.

Christie, P., Butler, D.,and Potterton, M. (2007). Report of ministerial committee: schools that work. October (pp. 1–138).

Christie, P. (2008). *Opening the doors of learning: Changing schools in South Africa* (pp. 1-235). Johannesburg: Heinemann.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, 41, 778–820.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45(3), 655–681.*

Cochran, W. G., and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics*, Series A, 417-446.

Coetzee, M. (2014). School quality and the performance of disadvantaged learners in South Africa (No. 22/2014). Working Papers. Stellenbosch University, Department of Economics.

Cohen, M and Seria, N. (2010). South Africa struggles to fix dysfunctional schools (Update2) - Bloomberg. Market Snapshot Bloomberg.

Connors, A. F., Speroff, T., Dawson, N. V, Thomas, C., Harrell, F. E., Wagner, D., Knaus, W. A. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *JAMA*, 276(11), 889–97.

Cook, T. D., Campbell, D. T., and Peracchio, L. (1990). Quasi Experimentation. In *Handbook of Industrial and Organizational Psychology* (pp. 491–576).

Creemers, B. P., and Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, 17(3), 347-366.

Crouch, L. A. (1996). Public education equity and efficiency in South Africa: Lessons for other countries. *Economics of Education Review*, 15(2), 125-137.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187–199.

Da Maia, C. C. P. (2012, December 1). Understanding poverty and inequality in Mozambique : the role of education and labour market status. Stellenbosch : Stellenbosch University.

Dauber, S. L., and Epstein, J. L. (1993). Parents' attitudes and practices of involvement in inner-city elementary and middle schools. *Families and schools in a pluralistic society*, 53-71.

DBE (Department of Basic Education). (2013). EMIS Ordinary Schools National Master List.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95, 158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, 42, 528–554.

Dee, T., and West, M. (2008). The non-cognitive returns to class size (No. 13994). NBER Working Papers. National Bureau of Economic Research.

Dehejia, R. H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.

Dehejia, R. H., and Wahba, S. (2002). Propensity score-matching methods For nonexperimental causal studies. *The Review of Economics and Statistics Association*, 84(1), 151–161.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Department of Basic Education. Norms and Standards for Educators (2000). South Africa.

Department of Basic Education. Revised National Curriculum Statement (2002). South Africa.

Desai, Z. (2001). Multilingualism in South Africa with particular reference to the role of African languages in education. *International Review of Education*, 47, 323–339.

Dieltiens, V., Chaka, T., and Mbokazi, S. (2007). Changing school practice: The role of democratic school governance. In B. Malcolm, C., Motala, E., Motala, S., Moyo, G., Pamapalis, J., Thaver (Ed.), *Democracy, Human Rights and Social Justice in Education* (pp. 12–22). Centre for Education Policy Development.

Dieltiens, V., and Meny-Gibert, S. (2012). In class? Poverty, social exclusion and school access in South Africa. *Journal of Education*, 55, 127-144.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. University of Michigan, Mimeograph.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica*, 64(5), 1001–44.

Duncan, K. C., and Sandy, J. (2007). Explaining the performance gap between public and private school students. *Eastern Economic Journal*, 33(2), 177–191.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–151.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802-813.

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, 66(1), 1–26.

Entwistle, N. J., Thompson, S., and Tait, H. (1992). *Guidelines for promoting effective learning in higher education*. Edinburgh: Centre for Research on Learning and Instruction, University of Edinburgh.

Eren, O., and Henderson, D. J. (2011). Are we wasting our children's time by giving them more homework? *Economics of Education Review*, 30, 950–961.

Fan, J., and Gijbels, I. (1995). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, 4(3), 213–227.

Filmer, D., Hasan, A., and Pritchett, L. (2006b). A millennium learning goal: Measuring real progress in education (No. 97). Center for Global Development.

Financial and Fiscal Commission. (2013). Submission for the 2013/14 Division of Revenue (pp. 42–63).

Financial and Fiscal Commission. (2014). Submission for the Division of Revenue 2015/2016.

Fiske, E. B., and Ladd, H. F. (2004*). Elusive equity: Education reform in post-apartheid South Africa* (p. 269). Brookings Institution Press.

Fleisch, B. (2008). *Primary education in crisis: Why South African school children underachieve in reading and mathematics* (p. 162). Juta and Company Ltd.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. *Handbook of Labor Economics*.

Fraser, N. (1998). *Social justice in the age of identity politics: Redistribution, recognition, participation* (No. FS I 98-108). WZB discussion paper.

Friedman, J. H. (2014). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.

Friedman, J. H., and Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22, 1365–1381.

Friedman, J. H., and Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2, 916–954.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.

Frölich, M. (2004). Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *The Review of Economics and Statistics*, 86(1), 77–90.

Fuller, B. (1986). Raising School Quality in Developing Countries: What Investments Boost Learning? World Bank Discussion Papers 2.

Gardeazabal, J., and Ugidos, A. (2004). More on identification in detailed wage decompositions. *Review of Economics and Statistics*, *86*(4), 1034-1036.

Goldhaber, D., and Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics*, 89(1), 134–150.

Goldschmid, B., and Goldschmid, M. L. (1976). Peer teaching in higher education: A review. *Higher Education*, 5, 9–33.

Greene, W. H. (2003). Econometric Analysis.

Gustafsson, M. (2007). Using the hierarchical linear model to understand school production in South Africa. *The South African Journal of Economics*, 75(1), 84–98.

Gustafsson, M., Berg, S. van der, Shepherd, D., and Burger, C. (2010). The costs of illiteracy in South Africa (No. 14/2010). Working Papers. Stellenbosch University, Department of Economics.

Gustafsson, M., and Taylor, S. (2013). Treating schools to a new administration. The impact of South Africa's 2005 provincial boundary changes on school performance (No. 28/2013). Working Papers. Stellenbosch University, Department of Economics.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2), 315–332.

Hahn, J., Todd, P., and Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69, 201–209.

Hallinger, P., and Murphy, J. F. (1986). The social context of effective schools. *American Journal of Education*, 94, 328.

Hanushek, E. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 351–388.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, 61, 280–288.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24(3), 1141–77.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, 19, 141–164.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., and Rivkin, S. G. (2005). The market for teacher quality (No. 11154). National Bureau of Economic Research Working Paper Series.

Hanushek, E. A., and Rivkin, S. G. (2006). School quality and the black-white achievement gap (No. 12651). National Bureau of Economic Research Working Paper Series (Vol. No. 12651).

Hanushek, E. A., and Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3), 607–668.

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models* (p. 352). CRC Press.

Hastie, T., and Tibshirani, R. (2000). Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3), 196–223.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161.

Heckman, J. J., Ichimura, H and Todd, P. E. (1998). Matching as an econometric estimator evaluation. *The Review of Economic Studies*, 65, 261–294.

Heckman, J. J., and Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, *30*(1), 239-267.

Heckman, J., and Jr, R. R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1), 239–267.

Heckman, J., and Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, 107–111.

Heckman, J. J., and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica*, *73*(3), 669-738.

Heneveld, W. (1994). Planning and monitoring the quality of primary education in Sub-Saharan Africa. AFTHR Technical Note No. 14.

Hijmans, R. J., Phillips, S., Leathwick, J., and Elith, J. (2011). Package " dismo ." October (p. 55).

Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2, 172–177.

Hill, H. C., Rowan, B., and Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42, 371–406.

Hill, J., and Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25(13), 2230–56.

Hirano, K., and Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2, 259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003b). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Ho, D., Imai, K., King, G., and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199–236.

Horn, P.M., and Jansen, A.I. (2009). An investigation into the impact of tutorials on the perofmrance of economics students. *South African Journal of Economics*, 77(1), 179–189.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2014). Party: A laboratory for recursive partytioning.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

Iacus, S. M., King, G., and Porro, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1–24.

Iacus, S. M., King, G., and Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493), 345–361.

Imbens, G., and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79, 933–959.

Imbens, G. W., and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–75.

Imbens, G. W., and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics,* 142, 615–635.

Jacobi, M. (1991). Mentoring and undergraduate academic success: A literature review. *Review of educational research*, *61*(4), 505-532.

Johnston, C., and James, R. (2000). An evaluation of collaborative problem solving for learning economics. *The Journal of Economic Education*, 31(1), 13-29.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of political Economy*, 410-442.

Kamper, G. (2008). A profile of effective leadership in some South African high-poverty schools. *South African Journal of Education*, 28(1), 1–18.

Kang, J. D. Y., and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523–539.

Keele, L. (2012). Observational studies with group level treatments: The case of catholic schools.

Kelley, A. C., and Swartz, C. (1975). Student-to-student tutoring in economics. Journal of Economic Education, 52-55.

Kingdon, G. (1996). The quality and efficiency of private and public education: A case-study of urban India. *Oxford Bulletin of Economics & Statistics*, 58, 57–82.

Kline, P. (2011). Oaxaca-Blinder as a reweighting estimator. *The American Economic Review*, *101*(3), 532-537.

Kreif, N., Grieve, R., Radice, R., and Sekhon, J. S. (2013). Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services and Outcomes Research Methodology*, 13(2-4), 174–202.

Krieg, J. M., and Storer, P. (2006). How much do students matter? Applying the Oaxaca decomposition to explain determinants of adequate yearly progress. *Contemporary Economic Policy*, *24*(4), 563-581.

Kuhn, M. (2008). caret Package. *Journal Of Statistical Software*, 28, 1–26.

Ladd, H. (2008). Teacher effects: What do we know? In G. Duncan & J. Spillane (Eds.), Teacher quality: Broadening and deepening the debate (pp. 3–26). Evanston, IL: Northwestern University.

Ladd, H. F., and Fiske, E. B. (2008). Handbook of research in education finance and policy. *Education Finance and Policy*, 3, 149–150.

Lam, D., Ardington, C., and Leibbrandt, M. (2011). Schooling as a lottery: Racial differences in school advancement in urban South Africa. *Journal of Development Economics*, 95, 121–136.

Lampa, E., Lind, L., Lind, P. M., and Bornefalk-Hermansson, A. (2014). The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. *Environmental Health: A Global Access Science Source*, 13, 57.

Lavy, V. (2010). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries (No. 16227). National Bureau of Economic Research Working Paper Series.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3), 337–46.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142, 675–697.

Lee, D. S., and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48 (2), 281-355.

Lee, R. E. (1987). Assessing retention program holding power effectiveness across smaller community colleges. *Journal of College Student Development*, 29(3), 223–27.

Lemieux, T. (2002). Decomposing changes in wage distributions: a unified approach. *Canadian Journal of Economics*, 35, 646–688.

Lemon, A. (1999). Shifting inequalities in South African schools: Some evidence from the Western Cape. *South African Geographical Journal*, 81(2), 96–105.

Letseka, M. and Maile, S. (2008). High university drop-out rates: a threat to South Africa's future? Retrieved from http://www.pan.org.za/node/8380

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2014). Balancing covariates via propensity score weighting. *arXiv preprint arXiv:1404.1785.*

Linnakyla, P., Malin, A., and Taube, K. (2004). Factors behind low reading literacy achievement. *Scandinavian Journal of Educational Research*, 48, 231–249.

Loh, W. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361–386.

Malcolm, C. (2000). Why Some "disadvantaged" Schools Succeed in Mathematics and Science: A Study of "feeder" Schools (p. 148). Mimeo.

Maxwell, M. (1990). Does tutoring help? A look at the literature. *Review of Research in Developmental Education*, 7(4), n4.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4), 403–25.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142, 698–714.

McEwan, P. J. (2008). Can schools reduce the indigenous test score gap? Evidence from Chile. *Journal of Development Studies*, 44, 1506–1530.

McVicar, D. (2001). School quality and staying-on in Northern Ireland - Resources, peer groups and ethos. *The Economic and Social Review*, 32, 131–151.

Mda, T. and Erasmus, J. (2008). Educators: scarce and critical skills research project. Retrieved from http://www.lmip.org.za/document/educators-scarce-and-critical-skills-research-project

Metzler, J., and Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics*, 99, 486–496.

Mincer, J. (1974). Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2. NBER (Vol. I).

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, 13, 125–145.

Mora, R. (2008). A nonparametric decomposition of the Mexican American average wage gap. *Journal of Applied Econometrics*, 23, 463–485.

Morgan, J. N., and Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association,* 58, 415–434.

Motala, S., & Pampallis, J. (2005). *Governance and finance in the South African schooling system: the first decade of democracy*. Centre for Education Policy Development.

Mouton, N., Louw, G., and Strydom, G. (2012). A historical analysis of the post-apartheid dispensation education in South Africa (1994-2011). Retrieved from http:// dspace.nwu.ac.za/handle/10394/10703

Mtika, P., and Gates, P. (2010). Developing learner-centered education among secondary trainee teachers in Malawi: The dilemma of appropriation and application. *International Journal of Educational Development*, 30(4), 396–404.

Mullens, J. E., Murnane, R. J., and Willett, J. B. (1996). The contribution of training and subject matter knowledge to teaching effectiveness: A multilevel analysis of longitudinal evidence from Belize. *Comparative Education Review*, 40(2), 139–157.

Munley, V. G., Garvey, E., and McConnell, M. J. (2010). The effectiveness of peer tutoring on student achievement at the university level. *American Economic Review*, 100, 277–282.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1), 141-142.

National Treasury. (2003). Intergovernmental Fiscal Review.

Ñopo, H. (2008). Matching as a tool to decompose wage gaps. *The Review of Economics and Statistics,* 90(2), 290-299.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14, 693–709.

OECD. (2008). Reviews of National Policies for Education - South Africa.

Pendlebury, S., and Enslin, P. (2004). Social justice and inclusion in education and politics: the South African case. *Journal of Education*, 34, 31–50.

Qin, X., and Han, J. (2008). Variable selection issues in tree-based regression models. *Transportation Research Record: Journal of the Transportation Research Board*, (2061), 30-38.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Reeves, C. A. (2005). The Effect of "opportunity-to-learn" and classroom pedagogy on mathematics achievement in schools serving low socio-economic status communities in the Cape Peninsula. University of Cape Town. Mimeo.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, 31, 172–181.

Ridgeway, G. (2007). Generalized boosted models : A guide to the gbm package. *Compute*, 1, 1–12.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73, 417–458.

Robeyns, I. (2006). The capability approach in practice. *Journal of Political Philosophy*, 14(3), 351–376.

Robeyns, I. (2009). The capability approach. In Handbook of Economics and Ethics (p. 39).

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429), 106–121.

Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *Journal of the American statistical Association*, 97(457), 183-192.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21(1), 57–71.

Rotnitzky, A., and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4), 805–820.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–48.

Rubin, D., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association, 95(450), 573–585.

Sakellariou, C. (2012). Decomposing the increase in TIMSS scores in Ghana: 2003-2007 (p. 62). World Bank.

Sandkull, O. (2005, August). Strengthening inclusive education by applying a rights-based approach to education programming. *In ISEC Conference, Glasgow* (pp. 1-9).

Schapire, R. E. (2003). The boosting approach to machine learning: an overview. *Nonlinear Estimation and Classification*, 171, 149–171.

Scheerens, J. (1999). School effectiveness in developed and developing countries: A review of the research evidence. World Bank, Human Development Network.

Schmidt, H. G., and Moust, J. H. (1995). What makes a tutor effective? A structural-equations modeling approach to learning in problem-based curricula. *Academic Medicine : Journal of the Association of American Medical College,* 70, 708–714.

Schneeweis, N. (2011). Educational institutions and the integration of migrants. *Journal of Population Economics*, 24, 1281–1308.

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a Stata plugin. *Stata Journal*, 5(3), 330.

Schultz, T. W. (1961). Investment in human capital. *American Economic Review*, 51(1), 1–17.

Schwerdt, G., and Wuppermann, A. C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review*, 30, 365–379.

Sela, R. J., and Simonoff, J. S. (2011). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2), 169–207.

Sela, Rebecca J. and Simonoff, J. . (2011). REEMtree: Regression trees with random effects. R package version 0.90.3.

Selod, H., and Zenou, Y. (2003). Private versus public schools in post-Apartheid South African cities: theory and policy implications. *Journal of Development Economics*, 71(2), 351–394.

Sen, A. (1997). Development and thinking at the beginning of the 21st Century. Vol, (2), 3–5.

Shalem, Y., Sapire, I., and Huntley, B. (2013). Mapping onto the mathematics curriculum – an opportunity for teachers to learn. *Pythagoras*, 34(1), 10 pages.

Shepherd, D. (2013). A question of efficiency: decomposing South African reading test scores using PIRLS 2006 (No. 19/2013). Working Papers. Stellenbosch University, Department of Economics.

Shepherd, D. L. (2011). Constraints to school effectiveness: what prevents poor schools from delivering results? (No. 13/2011). Working Papers. Stellenbosch University, Department of Economics.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57, 1–23.

Słoczyński, T. (2015). The Oaxaca–Blinder unexplained component as a treatment effects estimator. *Oxford Bulletin of Economics and Statistics* ,77(4), 588-604.

Sohn, K. (2012a). A new insight into the gender gap in math. *Bulletin of Economic Research*, 64(1), 135–155.

Sohn, K. (2012b). The dynamics of the evolution of the Black–White test score gap. *Education Economics,* 20(2), 175–188.

Spaull, N. (2011). A preliminary analysis of SACMEQ III South Africa (No. 09/2013). Working Papers. Stellenbosch University, Department of Economics.

Spaull, N. (2012). Education in SA: A tale of two systems. Retrieved June 15, 2015, from http://www.politicsweb.co.za/news-and-analysis/education-in-sa-a-tale-of-two-systems

Spaull, N. (2013). Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, 33(5), 436–447.

Spaull, N., and Kotze, J. (2014). Starting behind and staying behind in South Africa: The case of insurmountable learning deficits in mathematics (No. 23/2014). Working Papers. Stellenbosch University, Department of Economics.

Spreen, C.A., Vally, S. (2006). The Globalisation of education policy and practice in South Africa. In G. Martell (Ed.), *Education's Iron Cage and its Dismantling in the New Global Order*. Canadian Centre for Policy Alternatives, Toronto.

Staden, S. van, and Bosker, R. (2014). Factors that affect South African Reading Literacy Achievement: evidence from prePIRLS 2011. *South African Journal of Education*, 34(3), 1-9.

Strobl, C., and Boulesteix, A. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *Bioinformatrics*, 8(1), 25.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics,* 25(1), 1.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153), 65–66.

Subrahmanian, R. (2002). Citizenship and the "right to education": Perspectives from the Indian context. *IDS Bulletin*, 33(2), 1–10.

Tan, J.-P., Lane, J., and Coustere, P. (1997). Putting inputs to work in elementary schools: What can be done in the Philippines? *Economic Development and Cultural Change*, 45(4), 857–79.

Tansel, A. (1999). General versus vocational high schools and labor market outcomes in Turkey (No. 9905). SSRN Electronic Journal.

Taylor, N. (2008, February). What's wrong with South African schools. In *What's Working in School Development Conference, JET Education Services, Cape Town*.

Taylor, N., Muller, J., and Vinjevold, P. (2003). *Getting schools working: Research and systemic school reform in South Africa* (p. 151). Pearson South Africa.

Taylor, S. (2011). Uncovering indicators of effective school management in South Africa using the National School Effectiveness Study (No. 08/2011). Working Papers. Stellenbosch University, Department of Economics.

Taylor, S., and Coetzee, M. (2013). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach (No. 19/2013). Working Papers. Stellenbosch University, Department of Economics.

Taylor, S., and Yu, D. (2009). The importance of socio-economic status in determining educational achievement in South Africa (No. 07/2009). Working Papers. Stellenbosch University, Department of Economics.

Taylor, N. and Taylor, S. (2013). Teacher knowledge and professional habitus. In T. Taylor, N., van der Berg, S. and Mabogoane (Ed.), *Creating Effective Schools* (Pearson So.). Cape Town.

Taylor, Nick and Vinjevoild, P. (1999). Getting learning right: report of the President's Education Initiative Research Project.

Thistlethwaite, D. L., and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317.

Tikly, L. (2011). Towards a framework for researching the quality of education in low-income countries. *Comparative Education*, 47, 1–23.

Tikly, L., and Barrett, A. (2011). Social justice, capabilities and the quality of education in low income countries. *International Journal of Educational Development*, 31(1), 3–14.

Timaeus, I. M., and Boler, T. (2007). Father figures: The progress at school of orphans in South Africa. *Aids*, 21, S83-S93.

Timæus, I. M., Simelane, S., and Letsoalo, T. (2013). Poverty, race, and children's progress at school in South Africa. *Journal of Development Studies*, 49(2), 270–284.

Todd, P., and Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485), F3–F33.

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, 32, 321–345.

Traskin, M., and Small, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: A tree approach. *Statistics in Biosciences*, 3(1), 94–118.

Unterhalter, E. (2007). *Gender, schooling and global social justice*. Psychology Press.

Van der Berg, S. (2007). Apartheid's enduring legacy: Inequalities in education. *Journal of African Economies,* 16(5), 849–880.

Van der Berg, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation*, 34, 145–154.

Van der Berg, S., Wood, L., and le Roux, N. (2002). Differentiation in black education. *Development Southern Africa*, 19, 289–306.

Van der Berg, S., Girdwood, E., Shepherd, D.L., van Wyk, C., Kruger, J., Viljoen, J., Ezeobi, O. and Ntaka, P. (2014). The Impact of the Introduction of Grade R on Learning Outcomes. *Available at SSRN.*

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43, 1249–1287.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.

Walker, M. (2006). Towards a capability-based theory of social justice for education policy-making. *Journal of Education Policy,* 21, 163–185.

Watson, G. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics*, Series A.

Wayne, A. J., and Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73, 89–122.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8), 826–33.

Wilburn, S. (2013). How the "outside" becomes "inside": the social orientation of South African teachers' expectations for learning. *Journal of Education*, (58), 87–110.

Woo, M.-J., Reiter, J. P., and Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, 27, 3805–3816.

Woolman, S., and Fleisch, B. (2006). South Africa's unintended experiment in school choice: how the National Education Policy Act, the South Africa Schools Act and the Employment of Educators Act create the enabling conditions for quasi-markets in schools. *Education and the Law*, 18(1), 31–75.

Yamauchi, F. (2011). School quality, clustering and government subsidy in post-apartheid South Africa. *Economics of Education Review,* 30(1), 146–156.

Yates, C. (2007). Teacher education policy: International development discourses and the development of teacher education. In *Teacher Policy Forum for Sub-Saharan Africa* (pp. 6–9).

Yun, M. S. (2005). A simple solution to the identification problem in detailed wage decompositions. *Economic Inquiry*, 43, 766–772.

Yun, M. S. (2008). Identification problem and detailed Oaxaca decomposition : a general solution and inference. *Journal of Economic and Social Measurement*, 33(1), 27–38.