

https://dspace.org/

The contribution of Randomised Control Trials (RCTs) to improving education evaluations for policy: evidence from developing countries and South African case studies

Mohohlwane, Nompumelelo Lungile

2016

http://hdl.handle.net/10539/22682

Mohohlwane, Nompumelelo Lungile (2016) The contribution of Randomised Control Trials (RCTs) to improving education evaluations for policy: evidence from developing countries and South African case studies, University of the Witwatersrand, Johannesburg, <http://wiredspace.wits.ac.za/handle/10539/22682>

Downloaded from a DSpace repository. https://dspace.org/

THE CONTRIBUTION OF RANDOMISED CONTROL TRIALS (RCTs) TO IMPROVING EDUCATION EVALUATIONS FOR POLICY: EVIDENCE FROM DEVELOPING COUNTRIES AND SOUTH AFRICAN CASE STUDIES

Nompumelelo Lungile Mohohlwane

A research report submitted to the Wits School of Education, University of Witwatersrand, in partial fulfilment of the requirements for the degree Master of Education

Supervisors: Professor Brahm Fleisch
Dr Stephen Taylor

First Submitted March 2016
Revised Submission 17 October 2016

ABSTRACT

As access to formal schooling has expanded all over the world, there is acknowledgement that the quality of learning in many schooling systems, including South Africa, is extremely weak. Nationally representative samples of South African children participated in the PIRLS 2006 and pre-PIRLS 2011 studies, along with 48 other countries as a benchmarking exercise to measure the literacy levels of primary schools according to international standards. The PIRLS 2006 study indicated that more than 80% of South African children had not yet learned to read with meaning by grade 5. The pre-PIRLS results provided a new baseline of reading literacy levels for Grade 4 learners in South Africa, 29% of Grade 4 learners that participated did not have the rudimentary reading skills required at a Grade 2 level. Learners tested in African languages, particularly Sepedi and Tshivenda, achieved the lowest performance overall and were considered to be educationally at risk (University of Pretoria, 2012).

The context in which schooling takes place is key in understanding learner performance in South Africa. After decades of differential provision of education on the basis of race, the education system has been overhauled since the early 1990s. The South African government has introduced several initiatives and policies to address these systemic imbalances. All things considered, South Africa's learner performance has remained poor, even relative to several poorer countries in the region.

There is a wealth of research describing weaknesses in the education system. However, going a step further and identifying resources and practices that actually improve learner performance is central to improving education planning, policy and ultimately classroom practice. Rigorous evidence on classroom-based practice and resources that will have a measurable effect on learner performance in a developing country like South Africa is limited. The most significant shortfall of non-experimental evaluation methods (including qualitative and many quantitative approaches) is the absence of a valid estimate of the counterfactual – what outcomes would have been obtained amongst programme beneficiaries had they not received the programme. This often leads to the reporting of large positive effects of programmes being evaluated.

By using a lottery to allocate participants to an intervention and a control group, the Randomised Control Trial (RCT) methodology constructs a credible 'counterfactual' scenario – what might have happened to those who received an intervention had they not received it. This study provides a systematic literature-based argument on why RCTs should be part of the methodological options education researchers and policy makers consider in developing countries such as South Africa. Both the strengths and limitations of RCTs are discussed in light of the debate on RCTs and evaluation methods in education, as well as the technical critique of the methodology. The main critique of external validity is also elaborated on with efforts that may be taken to diminish the limitations discussed.

In addition, the study illustrates the value of RCTs using data from two South Africa RCTs on early grade reading interventions through a secondary analysis of the RCT data. The first case study in Chapter 4, is the Reading Catch-Up Programme (RCUP) conducted in Pinetown, KwaZulu-Natal. The main findings of the RCUP evaluation were that although learners in intervention schools improved their test scores between the baseline and the endline assessment, the learners in comparison schools improved by a similar margin. The results should contribute to a sobering realisation that the effects of the various interventions introduced by education stakeholders including NGOs and government are not obviously positive or more importantly, different from normal schooling. This points to the need to evaluate programmes before they are rolled out provincially or nationally, using RCTs and other rigorous methods.

The new analysis of data in this study explores the so-called "Matthew Effect" - the notion that initially better-performing children typically gain more from additional interventions and from schooling itself. The data from the RCUP RCT indicates that children with higher baseline test scores benefited from the intervention, whereas children with very low English proficiency at the outset did not benefit from the programme. Although females significantly outperform males in the reading tests used, there was no clear evidence of a differential effect of the

intervention by gender. The Matthew Effect therefore seems to be driven by prior knowledge and not gender or any other characteristic that was measured in the data.

The second case study in Chapter 5, is the Early Grade Reading Study (EGRS) conducted in the North West province. The EGRS may be seen as a more extensive follow-up to the RCUP to answer some of the unanswered questions. For example, will an early grade reading intervention that is implemented over a longer duration (two years) have an impact? Can intervening right at the start of school be a strategic point to intervene? Can a Home Language literacy intervention have lasting educational benefits?

In conclusion, although the policy formulation and evaluation process should draw on research using a variety of methods, the policy process will certainly be impoverished if there is a lack of research meeting two core criteria: interventions and findings that are relevant to the larger schooling population; and the precise measurement of the causal impact of interventions and/or policies. This study makes a clear literature-based argument on the contribution of internally valid methods, specifically RCTs in fulfilling these criteria and illustrates this with two case studies of RCTS. The study also provides a demonstration of the insights that are possible through secondary analysis founded on the richness of RCT data.

Key words: reading, education, language, South Africa, learner performance, education policy, randomised control trial

DECLARATION

I declare that this research report is my own unaided work completed under the supervision and guidance of my supervisors. It is submitted for the degree of Master of Education, at the WITS Schools of Education in the University of the Witwatersrand. It has not been submitted before for any degree or examination in any other university.

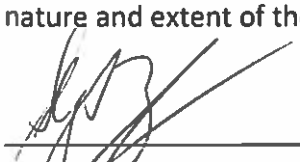
Notwithstanding the above, minor extract of the chapters one, four and five of the present work have been submitted for publication and subsequently accepted. The publication of this articles does not infringe upon my right to use this article in this dissertation or to publish the completed dissertation. The publication is listed in full below:

Mohohlwane,N. and Taylor,S. (2015). Using Impact Evaluation for Education Policy Innovations: The Case of Early Grade Literacy in South Africa. Evaluation Matters. A Quarterly Knowledge Publication on Development Evaluation. Second Quarter 2015: 99- 107 (Appendix A)

Declarations with respect to co-authoring

Nature of contribution	**Extent of contribution (%)**
Co-development of the structure of the article, co-writing of the article mostly through revisions and editing	50%

The undersigned co-author hereby confirm that the declaration above accurately reflects the nature and extent of the contributions of the candidate and the co-author.



Signature of co-author: Dr Stephen Taylor



Signature of candidate: Ms Nompumelelo Mohohlwane

Date: 15 March 2016

ACKNOWLEDGEMENTS

The completion of this degree is a significant milestone in my academic, professional and personal journey. I have confidence that the knowledge I attained, the research and analysis skills I developed and the insight I gained into the education research and policy literature have enhanced my contribution to the education sector. As I think about my colleagues in the education research field and various agencies including the DBE, I believe that the responsibility of public education is in the right hands. I am continuously inspired by the men and women who consistently fight to make public education meaningful for underprivileged learners, even at the cost of personal advancement. “We may be small in number but we are mighty in heart (author unknown).”

I am eternally indebted to my supervisors Professor Brahm Fleisch and Dr Stephen Taylor. The contributions you made, although different, were always complementary and expanded my repertoire. I am particularly grateful for the countless hours of detailed guidance, red track changes that changed to brown for the sake of my mental health, and the STATA tutorials by Stephen Taylor over the past year of working on this research. His patience, kindness and generosity have made a bigger impression than any amount of knowledge he shared with me, I hope to extend the same grace to others. I also have to acknowledge the encouragement I received from Brahm Fleisch, who saw the completion of this work as a fait accompli before I had even written up anything. Thank you for helping me tease out my voice in this work through the high octane session we had, I can see the golden thread in what I think is now a coherent contribution to the field.

Thank you to the National Research Foundation for the bursary I received, which was made possible by Brahm Fleisch. Thank you to the Zenex Foundation and the DBE for allowing me to use their data. Thank you to the researchers that conducted the Pinetown Reading Catch-up Programme: Prof Brahm Fleisch, Dr Stephen Taylor, Dr Thabo Mabogoane and Volker Schöer; and those that are still leading the research on the Early Grade Reading Study.

To my parents, Anna and Timothy Nyathi, thank you for prioritising education in our family; you always said that education was our inheritance and you were uncompromising in your sacrifices to afford us every opportunity to learn. I do not think I would have become who I am today without your love and guidance, I am forever in your debt.

To my husband, Katlego Mohohlwane, thank you for your support and interest in my studies, even as that has meant compromising the time we could spend together. Our many late night conversations have sharpened my thoughts and helped me to distil the essence of my convictions about public education. You often had the displeasure of helping me think through half-baked ideas and your patience and long suffering towards me continue to astound me; I am blessed to have you as my partner and my best friend.

To my brother, extended family, my friends, and Rooted Fellowship my sincerest gratitude goes to you for your love, support and encouragement during this time. Before I conclude, it would be an injustice if I did not single out my study buddies: Carol Nuga Deliwe, Glodean Thani and Nompumelelo Dammie. Thank you for all the ingenious ideas about the research, writing tips and my wellbeing, I would not have survived this degree without you.

Lastly, thank you to God that all thing work together for the good of those who love him and have been called according to his purpose.

TABLE OF CONTENTS

ABSTRACT 2
DECLARATION 5
ACKNOWLEDGEMENTS 6
APPENDICES OVERVIEW 10
LIST OF TABLES 11
LIST OF FIGURES 12
ABBREVIATIONS AND ACRONYMS 13
CHAPTER 1: INTRODUCTION AND BACKGROUND 14
1.1 Introduction 14
1.2 Background 15
1.3 Research questions 18
1.4 Rationale 19
1.5 Chapter Outline 22
CHAPTER 2: SYSTEMATIC LITERATURE REVIEW ON RANDOMISED CONTROL TRIALS 24
2.1 Introduction 24
2.2 Background 24
2.3 Overcoming the Problem of Selection Bias 26
2.4 Randomised Control Trials 29
2.5 Instrumental Variables 32
2.6 Regression Discontinuity Design 34
2.7 The Theory of Change and Binding Constraints 36
2.8 The Technical Requirements for Randomised Control Trials 43
2.8.1. Sampling: Sample Size, Statistical Significance and Confidence Intervals 43
2.8.2. Baseline Data Collection 50
2.8.3 Data analysis 51
2.9 Overcoming Publication Bias 52
2.10 Critiques and Caveats of RCTs 53
2.10.1 Epistemological Criticism 53
2.10.2 Caveats of Randomised Control Trials 59
CHAPTER 3: LITERATURE REVIEW 65
3.1 Introduction 65

3.2 International Literature on Experimental and Quasi-experimental Studies in Education 66
3.3 International Literature on Reading RCTs 70
3.4 National Literature on RCTs in Education 76
3.5 Conclusion of Literature Review 80
CHAPTER 4: CASE STUDY 1: READING CATCH –UP PROGRAMME 81
4.1 Introduction 81
4.2 Background 82
4.3 Intervention Design, Sample, and Data Collection 83
4.4 Learner Characteristics 86
4.5 Learner Performance 91
4.6 Conclusion 99
CHAPTER 5: CASE STUDY 2: EARLY GRADE READING STUDY 101
5.1 Background 101
5.2 Sampling and Data Collection 103
5.2.1 Baseline Data Collection 103
5.3 Learner Characteristics 104
5.4 Learner Performance 106
5.5 Conclusion 113
CHAPTER 6: CONCLUSION 115
6.1 Summary of Literature: Systematic Literature Review on the RCT Methodology 116
6.2 Summary of Literature: Literature Review on International and National Experimental and Quasi-experimental Research in Education 117
6.3 Summary of the Case Studies: RCUP and EGRS 119
6.4 Concluding Remarks 121
7. REFERENCES 123
APPENDIX A: Declaration - Article ..... **Error! Bookmark not defined.**
APPENDIX B: CHAPTER 4 RCUP Letter of Permission ..... **Error! Bookmark not defined.**
APPENDIX C: CHAPTER 4 RCUP Main Report ..... **Error! Bookmark not defined.**
APPENDIX D: CHAPTER 4 RCUP Pre-analysis Plan ..... **Error! Bookmark not defined.**
APPENDIX E: CHAPTER 4 EGRS Baseline Report ..... **Error! Bookmark not defined.**
APPENDIX F: CHAPTER 4 EGRS Letter of Permission ..... **Error! Bookmark not defined.**

APPENDICES OVERVIEW

APPENDIX A: Evaluation Matters Article.....130

The article discussed the South African low learner performance context, and the urgency of intervention to remedy this. The importance of evaluating, particularly using experimental and quasi-experimental methods is discussed. Two experimental studies, the Reading Catch-Up Programme(RCUP) and the Early Grade Reading Study(EGRS) conducted through a partnership between education researchers, government and donors, two are discussed.

APPENDIX B: RCUP Letter of Permission.....131

Letter of approval for the use of RCUP data for this research report.

APPENDIX C: RCUP Main Report.....132

The RCUP report is structured into four sections. The first is a brief introduction; followed by a detailed description of the study method focusing on a description of the intervention, the randomised control trial (RCT) methodology, the rationale for the selection of the study site, and the data-collection processes. The third section presents the major findings, including both information from a qualitative case study undertaken during the intervention and the results of the pre- and post-testing. The final section of the report considers the implications of the study

APPENDIX D: RCUP Pre-analysis Plan.....133

The RCUP Pre-analysis Plan provides the rationale for the study, description of the sample, Theory of Change, hypothesis to be tested including how this will be done through the provision of formulas and analysis descriptions for the main findings and heterogeneous effects.

APPENDIX E: 4 EGRS Baseline Report.....134

This report describes the results from the baseline data collection that was administered in 230 schools in February 2015. In each school a random sample of 20 grade 1 learners participated in oral assessments of reading and pre-reading skills. The results on the baseline test confirms the success of the randomization: on all measures of reading ability there is a good balance across the four treatment groups.

APPENDIX F: EGRS Letter of Permission.....135

Letter of approval for the use of EGRS data for this research report.

LIST OF TABLES

Table 1: Learner characteristics by gender and age…….. 86
Table 2: Learner characteristics by location and school quintile..... 87
Table 3: Learner performance by gender…… 90
Table 4: Decile average gain score for the Intervention Group and Comparison Group 93
Table 5: Main regressions…….. 94
Table 6: Regressions of overall test and subcategories with intervention and baseline interaction…….. 95
Table 7: Regressions of language subtest with intervention and performance categories 96
Table 8: Regressions of spelling subtest with intervention and performance categories 97
Table 9: Regressions of spelling subtest and performance categories with intervention and performance categories…….. 98
Table 10: Regressions of language subtest and performance categories with intervention by gender…….. 99
Table 11: Learner characteristics by gender and age……. 104
Table 12: Learner characteristics by location and school quintile..... 105
Table 13: Summary statistics on baseline learner performance 107
Table 14: Learner performance between the 10th and 90th percentile – letters correct and digit span…….. 111
Table 15: Regression of baseline score controlling for parental writing ability and gender .. 112

LIST OF FIGURES

Figure 1 National School Nutrition Programme Theory of Change 39
Figure 2: Random sampling and randomised assignment of treatment 60
Figure 3: Test score gains compared to cost-effectiveness in 30 RCTs 69
Figure 4: Baseline Performance by Intervention Group 88
Figure 5: Mean score for Intervention and Comparison Group for the baseline and endline test…….. 89
Figure 6: Average gain score for the Intervention Group and Comparison Group 92
Figure 7: Kernel density curves for Section B (letters correct) by treatment arm 109
Figure 8: Kernel density curves for Section C (Short-term memory) by treatment arm 110

ABBREVIATIONS AND ACRONYMS

ANA	Annual National Assessments
CRC	Community Reading Coach
DBE	Department of Basic Education
EGRA	Early Grade Reading Assessment
EGRS	Early Grade Reading Study
GPLMS	Gauteng Primary Language and Mathematics Strategy
HSRC	Human Sciences Research Council
IV	Instrumental Variable
NSC	National Senior Certificate
PED	Provincial Education Department
PIRLS	Progress in International Reading Literacy Study
RCT	Randomized Controlled Trial
RDD	Regression Discontinuity Design
SACMEQ	Southern and Eastern African Consortium for Monitoring Education Quality
SES	Socio-Economic Status
TIMSS	Trends in International Mathematics and Science Study

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Introduction

There is world-wide consensus on the importance of education and recognition that the significant contribution that sets it apart from other endeavours is the systematic development of knowledge and cognitive skills; in addition to its social, psychological and other benefits which will not be elaborated on in this study. As access to formal schooling has expanded all over the world, there is acknowledgement that the quality of learning in many schooling systems is extremely weak. Access to education is not in itself sufficient but rather serves as a necessary condition for education, with the emphasis of education being the attainment of cognitive skills and knowledge (Hanushek and Woessmann, 2007). A study by Taylor and Spaull (2015) based on an analysis of the performance of learners in 11 Southern and East African countries through a new measure that reflects both access to schooling and actual learning, 'Access-to-Learning', reflects on the issue of access and quality in sub-Saharan Africa. The study is a case in point in substantiating that access does not necessarily result in learning unless concerted efforts are made to deliver high quality education, particularly for the poorer learners within the system.

There is a wealth of research describing weaknesses in the education system however, identifying the most effective inputs for improving learner performance is central to improving education planning, policy and ultimately classroom practice. Rigorous evidence on classroom-based practice and resources that will have a measurable effect on learner performance in a developing country like South Africa is limited. Research that evaluates interventions is often based on case studies or small pilot studies. These methods are useful in their small-scale detail, but larger samples that will allow us to generalise to a larger population are required to inform policy decisions[1].

[1] Public policy is a statement of intent or an action plan to respond to a real or perceived problem in society. Decision-making about which areas or specific problems to address takes place amidst competing priorities. The factors considered include research and evaluations in addition to politics, economic cycles, electoral cycles and

This study contributes to the debate on effective evaluation methods in education research by reviewing Randomised Control Trials (RCTs) as one methodology that can be used in education research. The discussion focuses on the rationale, value and limitations of RCTs and also reflects on the criticism and critiques of this approach. In addition, the study seeks to illustrate the value of RCTs using data from two South Africa RCTs on early grade reading interventions.

1.2 Background

The performance of learners in South Africa, measured by several national and international assessments, has consistently been poor. The assessments referred to include the Annual National Assessments (ANAs), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and the study of the Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ).

Nationally representative samples of South African children participated in the PIRLS 2006 and pre-PIRLS 2011 studies, along with 48 other countries as a benchmarking exercise to measure the literacy levels at the primary school level according to international standards. The PIRLS 2006 study indicated that more than 80% of South African children had not yet learned to read with meaning by grade 5. In 2011 South Africa participated in the pre-PIRLS exercise along with Botswana and Colombia. The pre-PIRLS 2011 provided a new baseline of reading literacy levels for grade 4 learners in South Africa. The assessments were administered in all 11 official South African languages and designed to be less demanding than the main PIRLS assessment and yet 29% of grade 4 learners that participated in the Pre-PIRLS did not have the rudimentary reading skills required at a grade 2 level. Learners tested in African languages, particularly Sepedi and Tshivenda, achieved the lowest performance overall and were considered to be educationally at risk (University of Pretoria, 2012).

---

other interests. As such the availability of empirical information does not necessarily translate into policy decisions.

The assessments revealed that learners in the education system were performing well below the required levels, with a gap that spanned more than a year by the end of the Intermediate Phase in primary school, compared to other countries. The learning inequality gap within South Africa is similarly large, with learners in urban areas approximately two years ahead of those in rural or township areas (University of Pretoria, 2012).

Similar learner performance gaps have been reported in numeracy/mathematics in studies including TIMSS 2011, ANA and SACMEQ. South Africa participated in TIMSS at the grade 8/9 level along with 44 other countries. Although a substantial improvement in South Africa's grade 9 mathematics and science performance, especially at the bottom end of the achievement distribution, was observed between 2002 and 2011, these grade 9 learners were outperformed by grade 8 learners in nearly all other participating countries (Human Sciences Research Council, 2013). The poor performance in numeracy/mathematics is worth noting, although this study will only focus on literacy and reading. In summary, there is irrefutable evidence that the South African schooling system is not realising the critical goal of systematically developing knowledge and cognitive skills for a substantial proportion of the schooling population.

Literature on early learning emphasizes the importance of mastering certain learning foundations for the sake of all further learning. The literature refers to 'self-productivity', explaining that skills acquired during one period generally persist into the next period and may make the acquisition of other skills in another dimension easier (Girdwood, 2013). The low learner performance may therefore be seen as an indication of fundamentally weak reading foundations which thereafter have an increasingly negative effect on learners' ability to cope with curriculum requirements at higher grades.

The context in which schooling takes place is integral to understanding learner performance. As a result of the political and economic history of South Africa, education is seen as an instrument of change, and it is valued as a tool that has the potential to bring about equity in opportunity and outcomes. After decades of differential provisioning of education on the basis of race, the

education system has been overhauled since the early 1990s. This being said, the changes in the education system in the past 20 years have unfortunately had limited success in changing the nature of schooling. Several authors, such as Fleisch (2008) and Spaull (2014), argue that South African schooling still effectively has two systems in one. The first and largest part of the system consists of the historically disadvantaged schools and is characterised by inefficiency, including poor school management, continuous underperformance, and high and indiscriminate grade repetition and dropout. The second system refers to historically advantaged White and Indian schools, where learner performance is at a higher standard, parents make substantial fee contributions, organisational and instructional processes are more efficient, and schools are well endowed with resources and infrastructure.

Government efforts to create a single education system have included the establishment of a single department of education for all race groups, the reversal of unequal funding patterns, and substantial curriculum change. The main curriculum changes were the introduction of Outcome-Based Education (OBE) in 1997, followed by several revisions to the National Curriculum Statements (NCS) after widespread criticism of OBE, and finally the introduction of the repackaged National Curriculum Statement (NCS) often referred to as the Curriculum and Assessment Policy Statements (CAPS) curriculum in the last few years. The schooling system has now hopefully entered a phase of relative stability after 20 years of substantive curriculum changes.

Several initiatives and policies have also been introduced to address systemic resource imbalances. Tracking of progressive resourcing using school data indicates that since the early 1990s education spending has become increasingly well-targeted to poor schools (Gustafsson and Patel, 2008). Education received a 20% (R254 billion) share of the government consolidated expenditure in 2014, which is more than any other sector in South Africa (Department of Basic Education, 2014). Further initiatives include the introduction of no-fee schooling which is implemented in approximated 77% of public schools, scholar transport, and the provision of meals through the National School Nutrition Programme (NSNP) which fed 9 159 773 learners

in 21 400 primary, secondary and identified special schools in the poorest three official school poverty quintiles (Department of Basic Education, 2014).

Unfortunately, even with the political, resource and administrative developments, South Africa's learner performance has remained poor, even relative to several poorer countries in the region. The performance trends prompt an enquiry into the causes behind the limited ability of the education system to convert inputs into outcomes. The framework of national and provincial policies that support change in the system is firmly established, and there is a rich body of diagnostic research on system deficiencies, but the third critical component, empirical evidence-based policy interventions, is lagging behind. This situation makes it urgent and important to identify specific policies, programmes and large-scale interventions that can be shown to positively impact on learning outcomes.

1.3 Research questions

Although the policy formulation and evaluation process should draw on research using a variety of methods, the policy process will certainly be impoverished if there is a lack of research meeting the following two criteria: interventions and findings should be relevant to the larger school population; and the causal impact of interventions and/or policies should be measured precisely. This study will argue that RCTs can attain these criteria by the nature of their design and can thus make a substantial empirical contribution to policy-level intervention options.

This study focuses on RCTs by responding to the following questions:

- What is the value of RCTs as a methodological option for education research geared towards informing policy?
- How do case studies of RCTs illustrate the value of the methodology in gaining insights for policy and practice?

Scriven (1991) defines evaluation as determining the value or merit of programmes or policy options by identifying standards, performing empirical investigation using various techniques and integrating these into conclusions. In contrast, social research is restricted to empirical research based on observed measured or calculated data. Rogers (2014) categorises the relationship between research and evaluation into four approaches. The first, is to consider research and evaluation as two distinct and mutually exclusive categories; with the former focused towards the production of generalisable knowledge embedded in theory and the latter seen as interested in more specific and applied knowledge. An alternative approach is to view research and evaluation as two unrelated variables that are not mutually exclusive, noting that an activity can be both research and evaluation or neither. The third approach is to view evaluation as a subset of research based on the argument that all evaluation necessarily requires research but not all research requires evaluation. The fourth and final approach is to view research as a subset of evaluation based on the argument that research is one of the tasks involved in conducting evaluations.

In light of the definition by Scriven and the approaches proposed by Rogers, this study falls into the second approach-the view that an activity may be both research and evaluation. This contains features that directly relate to determining the value of RCTs in education policy fulfilling the evaluation criteria; as well as features of social research as the methodology applied and conclusions made in the case study component are based on empirical data rather. For this reason this study refers to both research and evaluation, noting the lack of mutual exclusivity in most instances for the purpose of the argument made.

1.4 Rationale

The debate on the most effective inputs and resource combinations for improving learner performance has been central to education planning for some time, forming a key area in research. The availability of standardised assessments has allowed us to measure the impact of schooling or enrolment in specific grades as well as providing a sense of what rates of

educational improvements are possible in developing countries. Research on production functions or resource combinations for optimal learner performance has recommended selected inputs with the caveat that effectiveness is dependent on other factors including equity, historical policy decisions and the overall developmental stage of the education system and the country in question (Boiseiere, 2004). Curriculum content and quality, as well as teachers are amongst the inputs identified as significant in research, as contributors to improving learner performance. These concepts may be unpacked to refer to appropriate subject selection, effective subject content transferral, and adequate time spent on quality teaching facilitated by the presence and expertise of quality teachers who are skilled in using different methods creatively and effectively to transfer knowledge.

There remains, however, a lack of rigorous evidence on effective classroom resources and practice that have a measurable effect on learner performance. This is particularly the case in developing countries, including South Africa. In cases where extensive research has been done, it is often focused on diagnosing areas requiring attention rather than evaluating possible solutions. Where interventions are evaluated it is often through conducting case studies or piloting in a small number of schools. There is great value in these methods including understanding school and classroom dynamics and identifying unexpected factors that may affect the success of interventions. A shortcoming of these approaches however, is that the implementation models used tend to be resource intensive and may be difficult to scale-up or replicate. Consequently, findings from such evaluations are very useful in providing detailed information on specific types of classroom but are difficult to apply to the broader schooling population.

The most significant shortfall of non-experimental evaluation methods (including qualitative and many quantitative approaches) is the absence of a valid estimate of the counterfactual – what outcomes would have been obtained amongst programme beneficiaries had they not received the programme. This often leads to the reporting of large positive effects of programmes being evaluated. Simply comparing recipients with non-recipients or pre-and post-

outcomes amongst recipients is usually not likely to provide a valid estimate of the counterfactual since recipients are usually systematically different to non-recipients and outcomes would change over time in any event. This is often the case in education research where education programmes are evaluated; similarly, much of the existing literature on large-scale early grade reading interventions in South Africa is open to this critique.

By using a lottery to allocate participants to an intervention and a control group, an RCT constructs a credible 'counterfactual' scenario – what might have happened to those who received an intervention had they not received it. With a large enough representative sample size to avoid the influence of outliers, random assignment, when done correctly using a lottery, will result in two groups who, prior to the intervention, can be assumed not to differ in any systematic way. Therefore, differences in the outcomes of the two groups after an intervention can be attributed to the causal impact of that intervention. Random assignment ensures unbiased estimates of causal impact; and with a large enough sample these effects can be precisely measured.

RCTs are less commonly used in developing countries than other research methods, but their use is increasing. There is a growing body of work on their rationale, appropriateness and methodology in the social sciences, including education. This study provides a systematic literature-based argument on why RCTs should be part of the methodological options education researchers and policymakers consider in developing countries such as South Africa. In addition to elaborating on how RCTs provide unbiased estimates of the impact of interventions in a manner that is generalisable to a larger population and therefore is relevant for policy, the study illustrates the insights that are possible through secondary research based on the richness of RCT data, demonstrating the value of RCTs through the analysis of two South African case studies by combining some reanalysis and replications from the studies with new analysis.

There is a distinct gap in the South African literature on the use of quasi-experimental and experimental methods including RCTs on early grade reading research or education more broadly. Those that have been identified in a review of the literature do not always appear to fulfil all the criteria for RCTs, this is elaborated on in the study. The two case studies selected to showcase RCTs are amongst the only studies that meet the criteria of RCTs fully, have been undertaken in the early grades and have a sample representative of the larger schooling population which is a requirement for policy research.

1.5 Chapter Outline

The chapters that follow respond to the research questions stated above, namely, what is the value of RCTs as a methodological option for education research geared towards informing policy; and how do case studies of RCTs illustrate the value of the methodology in gaining insights for policy and practice?

Chapter Two provides a systematic literature-based argument to establish the need to evaluate education interventions, and the problem of selection bias that occurs as part of the evaluation context. The contribution of quantitative methods towards eliminating selection bias is discussed starting with RCTs and the appropriate policy opportunities to introduce them. This is followed by a discussion on other internally valid methods such as Instrumental Variables, and the Regression Discontinuity Design, including examples of appropriate policy scenarios and the limitations of these methods. As this study focuses on the contribution of RCTs for policy-level education interventions, the remainder of the chapter focuses on the technical components of RCTs, including sampling, statistical significance, data collection and data analysis. The final section focuses on the critiques of RCTs as well as mechanisms to overcome the areas of weakness identified.

Chapter Three provides a summarised literature review of RCTs in education drawing on international and local literature. This includes international and local RCTs specifically focused

on reading, with an emphasis on early grade reading. Insights from the literature are discussed, and a critique of the studies is provided.

Chapter Four and Five illustrate the value of RCTs through a secondary data analysis of two South African RCT case studies. The first case study in Chapter Four, is the Reading Catch-Up Programme (RCUP) conducted in Pinetown, KwaZulu-Natal. The second case study in Chapter Five, is the Early Grade Reading Study (EGRS) conducted in the North West province. The sample and data collection process as well as the reliability of these case studies are discussed with the main focus of the chapter being an illustration of the kinds of research and policy insights that emerge from RCT data. The data analysis explores interactions between learner characteristics, including gender, and the extent of the Matthew Effect (the notion that initially better-performing children typically gain more from additional interventions and from schooling itself). The chapter concludes with an analysis of the second case study, measuring learner inequalities at the start of schooling and learner characteristics associated with inequalities.

Chapter Six concludes the study by summarising the findings and drawing out policy insights and implications and the contribution made to the literature on early grade reading research.

CHAPTER 2: SYSTEMATIC LITERATURE REVIEW ON RANDOMISED CONTROL TRIALS

2.1 Introduction

RCTs have played an important role in policy development in industrialised countries and increasingly in developing countries. This has been primarily in medicine through clinical trials of new treatments or drugs, and increasingly in agriculture, micro-lending and credit as well as education. In comparison to other forms of research including quantitative research or desktop reviews, RCTs are less common in developing countries; nevertheless there is a growing body of work on the rationale, appropriateness and application of the methodology in the social sciences including education. This chapter argues for the use of RCTs in education research for policy through a systematic examination of the literature on selected quantitative research methods relevant for the education context.

2.2 Background

A detailed review of South African education policy development, monitoring and evaluation is beyond the scope of this study. What is clear, however, is that despite many policy changes and new programmes, little is known about the ultimate impact of these initiatives on learning outcomes. The lack of a focus on impact evaluation is not unique to South Africa, as the following quote illustrates:

> "Development programs and policies are typically designed to change outcomes, for example, to raise incomes, to improve learning, or to reduce illness. Whether or not these changes are actually achieved is a crucial public policy question but one that is not often examined. More commonly, program managers and policy makers focus on controlling and measuring the inputs and immediate outputs of a program—how much money is spent, how many textbooks are distributed—rather than on assessing whether programs have achieved their intended goals of improving well-being" (World Bank, 2010, p.2).

In cases where large scale quantitative research is done – at least in South Africa – it is typically focused on diagnosing areas requiring attention rather than evaluating possible solutions. Alternatively, correlations and conditional correlations pointing tentatively to a possible causal relationship are presented using methods such as the education production function as seen in the work of Gustafsson (2007), Spaull (2012) and Van der Berg (2008). The limitation of these studies emanates from the fact that they rely on large sample surveys, which provide "observational" data in the sense that one can observe how particular outcomes of interest are associated with observed characteristics of learners, teachers and schools. So, for example, one might observe that schools with libraries have better-performing learners on mathematics tests, but that is not necessarily indicative of a causal relationship.

Where interventions are evaluated it is often through conducting a small number of case studies or piloting the intervention in a small number of schools (for example, Pretorius, 2014). The shortcoming of this approach is that the implementation model used in case studies or small-scale pilots is often resource intensive and may be difficult to replicate at a larger scale. These evaluation methods, although detailed in understanding how change takes place, do not use a large enough sample to permit precise inference of their findings to a larger population.

A focus on evaluation is emerging within the South African government and one of the mechanisms developed to institutionalise the practice, was the introduction of the National Evaluation Policy Framework in 2011. This policy framework includes a National Evaluation Plan (NEP) which commissions independent evaluations of priority government programmes in a partnership between the custodian department and the Department of Planning, Monitoring and Evaluation (DPME, 2014). Several Department of Basic Education (DBE) programmes have been evaluated through the NEP, namely the Grade R Programme, the Funza Lushaka Bursary Programme and the National School Nutrition Programme.

The evaluations referred to above are all retrospective evaluations, assessing how well programmes were implemented or if the intended programme goals were attained. Prospective impact evaluations, where programmes are evaluated prior to being taken to scale, remain extremely rare. One exception to this is the impact evaluation of a new set of study guides introduced by the DBE in 2012 (Department of Basic Education, 2013). There remains a clear gap in education literature on rigorous evidence on classroom-based practice and resources that will have a measurable effect on learner performance in a developing country like South Africa even within the focus of government-led research.

2.3 Overcoming the Problem of Selection Bias

The main purpose of evaluations is to identify whether a policy, programme or intervention, otherwise referred to as a 'treatment' in this study, made a difference or had the intended effect; this may be referred to as impact. However, the question of impact is more complex than establishing if a programme has been implemented or how well implementation has been executed. The major challenge in impact evaluation is the need to identify a counterfactual, what would have happened to programme recipients in the absence of the intervention? Since one can never actually observe a counterfactual in reality, one needs to use a "control group" or "comparison group" to try and provide a valid estimate of the counterfactual.

Before discussing the idea of a counterfactual further, it is worth noting, firstly, that a comparison of the pre- and post-outcomes amongst recipients of a programme as a method for measuring impact, is problematic. When conducting a simple "before and after" evaluation one has to assume that outcomes would not have changed in the absence of the intervention. This assumption is usually implausible and thus drawing causal conclusions from this kind of evaluation may not be useful. An example of this is simply comparing learner results before and after the addition of a library to a school and then attributing the change in results to the library. Although the library may have had an impact, there are numerous other factors which would also have affected learning outcomes in a school over the intervention period. Individual

learners will always be learning over time. Similarly, school performance tends to fluctuate over time for any number of reasons.

A second rather simplistic way to estimate the counterfactual is to compare outcomes amongst programmes beneficiaries to those amongst non-beneficiaries. This is sometimes referred to as the "simple difference" method. However, in the vast majority of real world programmes there is some systematic pattern driving who will be programme recipients. For example, we cannot compare learning outcomes of South African students who benefit from the National School Nutrition Programme (NSNP) to those who do not, since the programme explicitly targets schools in poor communities. We would no doubt observed better learning outcomes amongst those not receiving the programme but this would not represent a negative causal impact of the programme.

A substantial improvement on the pre- and post-method and the simple difference method is to combine the two to measure the so-called "difference-in-differences". If the change in performance over time is greater for programme beneficiaries than it is for non-beneficiaries, this could be interpreted as evidence of programme impact. However, this method still has to assume that the two groups would have been on parallel outcome trends in the absence of the intervention.

A fourth set of solutions to the problem of selection bias is to somehow account for differences between programme recipients and non-recipients. Suppose a minority of female learners from poor households manage to attend high-performing girls-only private schools. We may be interested to measure the impact of this. Of course, we cannot literally observe the same individual at the same time, in a poor performing school. We could however, identify a comparison group of girls coming from similarly poor households who attend less-resourced schools and compare learning outcomes. This method thus takes account of the Socio-Economic Status (SES) of girls in the two groups of schools. Multivariate regression methods as well as matching techniques, such as propensity score matching, fall into this broad category of

methods in that they take account of the observable characteristics of beneficiaries and non-beneficiaries. These methods can go a long way towards a credible estimate of programme impact, depending on the specific research setting. However, there may still be important unobservable characteristics which simultaneously determine programme assignment and outcomes. For example, the girls from poor households who attend high-performing schools may have more motivated parents who place a greater value on education than their equally poor counterparts attending weaker schools. It might not be possible to measure these factors, hence they are unobservable. These unobservable difference could however be driving the impact seen.

In summary, simply comparing recipients with non-recipients is usually not likely to provide a valid estimate of the counterfactual when programme assignment happens on the basis of some systematic non-random process. Individuals selecting themselves into a programme may be more motivated, more educated or wealthier than those who do not. Both the observable and unobservable characteristics of the treatment group, may affect outcomes. As such, careful matching of both is required for fair comparison.

Government targeting of programme or intervention recipients also means that recipients are often selected in a non-random way based on perceived needs such as poverty, low learner performance, geographical area or gender to mention a few. This means that, typically, recipients and non-recipients of interventions are not comparable. A South African example of this is the NSNP discussed above, which is a DBE intervention whereby learners are provided with a nutritious meal on every school day. The NSNP is targeted at learners in quintiles 1-3 based on National Treasury classifications of school SES with the aim of improving attendance, addressing short-term hunger and contributing towards improved concentration in class, and as a result, improved learner performance (DBE, 2010). The profile of these learners is characterised by a food-insecurity background and generally poor school performance, while non-recipients consisting of learners in quintiles 4 and 5 schools or independent schools, generally perform better and come from wealthier backgrounds. It follows then that it is not

plausible to estimate the impact of the NSNP on learning outcomes by comparing these systematically different groups. This is often the environment in which government interventions are rolled out and as such impact evaluation becomes a challenge.

All of the scenarios discussed above create what is known as selection bias in impact evaluations. There are various experimental and quasi-experimental quantitative methods for establishing internal validity and avoiding selection bias, these include Randomised Control Trials, the Regression Discontinuity Design and Instrumental Variables.

2.4 Randomised Control Trials

While various quantitative impact evaluation methods are available, the cleanest method for eliminating selection bias and identifying an internally valid estimate of the counterfactual is obtained through conducting a Randomised Controlled Trial (RCT). Through using a lottery to allocate participants to intervention and control groups, an RCT constructs a credible "counterfactual" scenario – what would have happened to those who received an intervention had they not received that intervention. With a large enough sample size to avoid the influence of outliers; random assignment will result in two groups who, prior to the intervention, can be assumed to be no different in any systematic way. This assumption of statistical equivalence can be tested through balance tests using observable characteristics, such as a baseline test score. Importantly, even unobservable characteristics, such as people's motivation, can be assumed to be no different across treatment groups. Therefore, any differences between the treatment group and the control group that are observed after the intervention can confidently be attributed to the causal impact of the intervention. Randomisation takes account of both observable and unobservable characteristics as all participants in the population of interest have an equal chance of being allocated to either the treatment or control group. There is therefore no selection bias when randomisation is done correctly and as such the method is internally valid (Duflo, Glennerster and Kremer ,2006).

Internal validity means there is no selection bias; high internal validity means we are able to place high confidence in the cause and effect relationship in our experiment, referred to as the average treatment effect. It is a product of the process of randomised assignment of treatment and control groups and reflects the ability to control for issues that would affect the causal interpretation of the treatment impact. Through the use of a representative sample and randomisation, it is highly plausible to expect that confounding variables that form part of the characteristics of the sample are equally distributed between the intervention group/s and the counterfactual group thus supporting the assertion of internal validity. In other words, internal validity is the assurance that the comparison group represents a true counterfactual and, as a consequence, the experiment is estimating the true impact of the intervention or programme (Gertler, Martinez, Premand, Rawlings and Vermeersch, 2010).

The randomisation referred to above is based on classic randomisation at the inception of an experiment. A variant to this is introducing randomisation into an existing programme. Opportunities may exist in circumstance where there may be oversubscription; where a programme is being amended and there is an opportunity for a phased-in approach; or where within-group randomisation may be introduced. These three variations of randomisation are often applicable to the manner in which government introduces programmes as there are often cost and capacity constraints that require a selection of participants. These methods of introducing randomisation may provide an opportunity for responding to implementation challenges in a fair, unbiased manner. The three variations are discussed below, with scenarios and examples.

The case of oversubscription (where there are more potential beneficiaries than the programme can cover) presents an opportunity for randomisation of allocation. The oversubscription may be a result of budget restrictions, and in such a case, random allocation of the treatment or intervention may be a fair way of selecting programme participants. Randomisation would occur after identifying those eligible for participation. In cases where limiting the recipients of an intervention is unacceptable but the budgetary and

oversubscription constraints still exist, the phased-in approach may be used. Two groups would still be created, with allocation into either group being randomised. One of these groups would then be allocated as the first recipients and the remaining group would initially serve as a comparison group and then later receive the same intervention. The phased-in approach may also encourage the comparison group to continue to participate as they are assured of the receipt of the intervention at the next phase of roll-out.

It is important to note that, with a phased-in approach, a sufficient time lag between phases is required to ensure an adequate opportunity to measure the effect of the intervention. However, the expectation to receive the intervention may affect the behaviour of eligible participants and thus bias the results by creating the conditions for an overestimated effect size of the intervention. A further limitation to the phased-in approach is that it limits the opportunity to evaluate long-term effects. Following well-defined cohorts that may no longer be eligible for reasons unrelated to the intervention may however, continue to provide a plausible control group against which to continue to measure longer-term effects of the intervention (Duflo et al., 2006). Governments and organisations in developing countries particularly, are often faced with the contexts referred to above, and this type of randomisation provides a possible solution to reducing bias in providing intervention or services.

There may be circumstances where the population of interest refuses to participate in a study unless they benefit directly. For example, a school may refuse to allow researchers to collect learner health or performance data if they are assigned to a comparison group and do not derive short-term benefits from an intervention. Within-group randomisation, which is randomly selecting children for participation within a school rather than randomly selecting intact schools, may be a mechanism to address this. Approximately 80% of South African public schools are part of the NSNP as mentioned above. In an RCT on the effectiveness of providing supplements in addition to the meal served daily, through within-group randomisation, learners in a different grade in each school could be randomly selected to be recipients of these supplements. As school managers know that some learners would benefit, even when this is

restricted to selected learners, they may be more amenable to participation. As allocation takes place in a different grade, an intervention and comparison group could still be maintained across schools. Within-group randomisation will only be suitable in a setting where spill over of programme benefits to non-beneficiaries within the group is unlikely to occur.

2.5 Instrumental Variables

A quasi-experimental method that addresses selection bias and yields high internal validity when convincingly applied is Instrumental Variables (IV), an important tool used in econometrics. IVs refer to when participation in a programme can be predicted by an incidental factor or 'instrumental' variable that itself would not be expected to be correlated with the outcome of interest (Glennerster and Takavarasha, 2013). Instrumental variables are sometimes referred to as natural experiments as they entail exploiting situations where "the forces of nature or government policy have conspired to produce an environment somewhat akin to a randomized experiment" (Angrist and Krueger,2001, p. 75). These 'natural experiments' may be enabled by occurrences that happen differently because of geographical location or environmental occurrences. The use of such geographical features as good instrument is demonstrated in the literature (Glennerster and Takavarasha, 2013:42). "Imagine a remote valley in Indonesia that is separated from a less remote, more densely populated area by a high ridge of mountains. In some parts of the valley, the population can get television (TV) reception designed for the more populated area on the other side of the mountain ridge… Whether a given community receives TV reception therefore, is uncorrelated with remoteness or land quality. The height of the nearby ridge affects the population only through its effect on their ability to receive TV signal. We can then use the height of the ridge to assess the impact of TV on outcomes such as social capital or attitudes towards women".

Owning a TV is a choice and simply comparing those that own TVs to those that do not in an effort to evaluate the impact of TV would result in selection bias. However, as in this example, if TV ownership is predicted by something that is not a choice such as the height of the nearest

mountain, this would sufficiently eliminate selection bias in evaluating the effect of TV ownership on perceptions or beliefs (Glennerster and Takavarasha, 2013). Those who own TVs may have had exposure to specific messaging on human rights, gender equity or different religions through various TV programmes and could thus be expected to have somewhat different views than their counterparts who have had no exposure.

The IV method allows for consistent estimation of causal relationships when the explanatory variables or covariates, are correlated with the error terms of a regression relationship. An IV is a variable that itself does not belong in the explanatory variables. IVs are used to address three main problems: omitted variable bias, measurement error in explanatory variables and reverse causality or confounding factors. IVs are often used when controlled experiments are not available or feasible. The credibility of the estimates of impact relies on the selection of suitable instruments: when an instrument is weak there is a resultant weak relationship between the instrument and the endogenous regressor, defined as a variable that is correlated with the error term. A good instrument is correlated with the endogenous regressor for reasons that may be explained and verified but uncorrelated with the outcome variable for reasons beyond its effect on the endogenous regressor. A good instrument is often difficult to determine, with success tied to detailed knowledge of the economic mechanisms and institutions determining the regressor of interest (Angrist and Krueger, 2001).

When implementing an intervention that is available to the full population of potential beneficiaries but only partially taken up, we cannot simply compare those who take up the intervention with those who do not, since their unobservable characteristics may differ systematically. However, we could encourage take-up by providing a small incentive to a randomly chosen group of potential beneficiaries. This is a special case of IVs known as the Encouragement Design. The description of the design refers to the intention to increase the chances that an intervention is received rather than the intervention itself. The encouragement design is relevant where randomisation of an intervention is not possible, so instead of randomising the treatment, there is random assignment of encouragement to participants for

take-up. This encouragement may be used as an IV for the take-up of the programme (Khandker, Kodwa and Samad, 2010). For example, we may have a Theory of Change stating that increased parental attendance at parent-teacher meetings improves learning outcomes within the school, through *inter alia* strengthened local accountability. Suppose a raffle with prize money was introduced at parent-teacher meetings within a randomly chosen group of schools. If we observed both improved parent attendance and improved learning outcomes within the group of schools which introduced the raffle, it would be valid to attribute the improved learning outcomes to attending parent-teacher meetings. This is because the only way that the raffle could have influenced learning outcomes is through attendance at parent-teacher meetings.

2.6 Regression Discontinuity Design

There are instances where a policy rule or programme allocation rule causes some people to receive a programme and others not to, based on certain conditions such as an administrative decision or limited resources. The policy cut-off point is normally clear and determined by policy makers or administrators rather than a discretionary choice by the targeted participants. If such a policy allocation cut-off occurs at an arbitrary point along a continuous dimension, such that those falling just above the cut-off are negligibly different from those falling just below the cut-off, then one might regard assignment as effectively random amongst those just either side of the cut-off. Of course, one cannot compare all those falling beneath the cut-off with all those above the cut-off, but just either side of the cut-off, the two groups may be comparable. This creates a type of localised RCT which may be categorised as a Regression Discontinuity Design (RDDs). This provides a quantitative method for estimating the causal effect of the intervention on participants (Imbens and Lemieux, 2008). The RDD enables the creation of a binary treatment variable of recipients and non-recipients around the cut-off point, effectively creating an intervention group and an internally valid counterfactual.

One way to further strengthen the RDD design is through matching of participants in the 'treatment group and the 'comparison group' around the cut-off point as defined above. Each participant is matched with at least one non-participant based on observable characteristics such as gender, age and school SES. This could be done through propensity score matching, where characteristics such as age and school SES are used to predict participation. Each participant is given a probability or propensity score for participation, and participants and non-participants are matched based on their propensity score (Glennerster and Takavarasha, 2013).

There are two main kinds of RDDs, the sharp and the fuzzy design. The sharp RDD occurs when compliance with the allocation cut-off is complete. I.e. everyone on one side of the cut-off receives the intervention and everyone on the other side does not receive the intervention. In this case one simply compares outcomes between the two groups on either side of the cut-off point.

In the Fuzzy RDD, compliance with the allocation rule is not perfect. There may be some who qualify for the intervention but choose not to participate and others who do not actually qualify but still manage to participate. As long as there is a discontinuous jump in the probability of participation at the relevant threshold the causal impact can still be estimated using an IV approach. In effect the fuzzy RDD is a special case of IV, in which the policy rule applying at the cut-off becomes an IV. If one observes both an increase in the probability of participation at the cut-off and an improvement in outcomes, then that improvement can be attributed to the programme.

An example for the potential use of an RDD in South Africa would be a policy rule that allocates an intervention to a subset of the population who fall within a certain threshold such as schools performing below 40% in the ANAs. Clearly one cannot simply compare all schools scoring less than 40% to all schools scoring more than 40%. However schools scoring 39% are arguably, essentially the same as schools scoring 41%. Therefore, within a narrow range around the cut-off point there is an internally valid intervention group and a plausible counterfactual group.

The RDD assumes that there would be a continuous relationship between the assignment variable and the outcome variable in the absence of the intervention. Therefore if a discontinuous relationship is observed at the point of the cut-off, it is assumed that this is representing the causal relationship of the intervention. If there is any other possible reason to expect a discontinuity at the cut-off point an RDD is invalid (Taylor,2014). Given this, RDDs are not often used in developing countries, for two main reasons, firstly, adherence to the eligibility criteria is often not strictly applied by officials rolling out the intervention, and secondly, officials generally have the latitude to use their discretion for what may be valid reasons, and this affects the plausibility of avoiding selection bias (Duflo et al., 2006). A further limitation of the RDD is that it provides what is called a local treatment effect; that is, it only provides information on the impact of the intervention for those around the cut-off point; it does not provide information about what the impact of the programme is for those far above or far below the cut-off. Bearing all this in mind, RDDs still present an opportunity for quantitative research that should be explored and considered depending on the context. There are several government programmes that are rolled out in a manner that could quite easily be slightly adjusted to facilitate an RDD evaluation in order to contribute, even in a limited capacity, to understanding programme or policy impact.

2.7 The Theory of Change and Binding Constraints

"Theories' of change are the ideas and beliefs people have – consciously or not – about why and how the world and people change" Hivos (2015). In the complex environment of governance, change is driven by various factors including political, cultural and societal interests. As such Theories of Change are built on the premise that social change is complex and dynamic. A Theory of Change (TOC) provides a strategic approach to understanding the programme or intervention design as perceived by various stakeholders. It may be used as an opportunity to engage critically by questioning all of the mechanisms, relationships and assumptions embedded within programmes and interventions; and may provide transparency

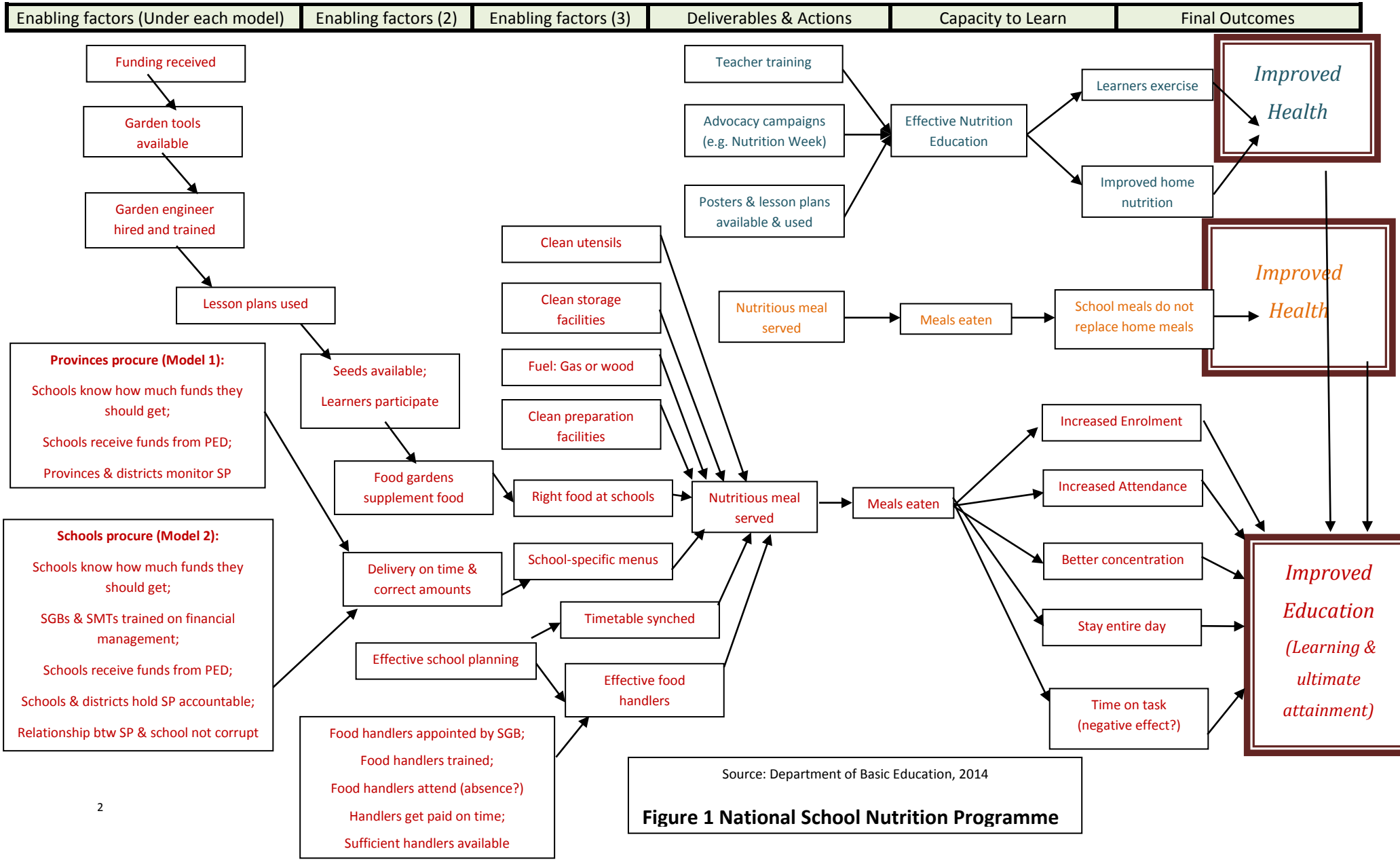
in programme design. A TOC articulates the causal relationship supporting the hypothesis behind programmes or interventions by describing how the causal mechanisms between activities and outputs will result in the anticipated outcomes (DPME, 2016). It is intended to be more than a list of activities and possible outcomes by providing the links between expected outcomes, the steps required to realise those outcomes and the assumptions underpinning the logic. A TOC may assist programme managers in bridging the gap between ambitious policies or programmes and the resources or structures available to support the implementation of these. A TOC also ensures a common understanding of how interventions or policies will work and thus improve coordination and collaboration trough this common understanding.

In the designing of interventions or programmes, particularly for policy purposes, there are often multiple stakeholders and an extended timeframe in which intervention or programme will be implemented. During the lifespan of the intervention or project there may be changes at various managerial levels including changes in the Ministry, Administrative leadership and even within programmes. It is therefore imperative to have an explicitly articulated TOC at the inception of programmes or interventions for sustainability and knowledge management in addition to the benefits already cited.

The core components of a TOC are establishing the desired outcomes as well as the underlying reason for this goal; analysing the current context; mapping the various plausible pathways for change; articulating the assumptions underpinning the pathways; determining the most strategic pathway and then establishing indicators and data that will be used to measure accomplishment of the outcomes Hivos (2015). As a TOC is a living document, it is important to continue to revisit and reshape it according to changing inputs, outputs, assumptions and even outcomes.

An example of a TOC based on the NSNP follows below (Department of Basic Education, 2015). Intervention development always has a context, goal and activities but this is often implicit with no clear documented TOC and careful consideration of the relationship between inputs,

activities, outputs and outcomes or impact. This is a limitation as without the careful consideration of these building blocks, well intended programmes may fail. In order to reach the intended programme outcome there is a link between the existence and execution of various enabling factors. Defining these relationships, testing the validity of the underlying assumptions and monitoring them are critical practices required to realise any programme. In the NSNP TOC it is clear that the main intended outcome is improved educational attainment but part of the main key activities enabling that are the preparation of a nutritious meal served within a specific timeframe in order to impact on classroom learning ability. The main assumptions underlying this, are that schools know what constitutes a nutritious meal and are able to replicate this across the provinces; and that they understand the need to serve the meal early to enable learning in class. The reality however, is that meals may be cooked to taste without nutritional value being carefully considered and may be served too late in the day to have the intended in-classroom benefit. Among the main contributing factors for the deviations from the intended implementation is often the lack of a clear documented TOC that has been well-communicated across the sector. Other factors include the complexities of the delivery of education and the set of variables contributing to improved learner attainment over and above nutrition, as well as the massive operational arrangements required to deliver the NSNP.



Source: Department of Basic Education, 2014

Figure 1 National School Nutrition Programme

[2] The abbreviations in the TOC refer to the various stakeholders involved, School Governing Body (SGB), School Management Team (SMT), Provincial Education Department (PED), Service Provider (SP).

Once a detailed TOC has been set out, collection of data on intermediate outcomes is an important method of measuring the validity of the TOC and the estimated programme impact. For example, a TOC regarding the impact of providing textbooks on learning must at some point assume that textbooks are actually used. Collecting data on textbook usage will be helpful when interpreting whatever impact on learning is observed. Collecting information on intermediate outcomes is also useful for understanding what is referred to as the "partial derivative" estimate of impact and the "total derivative" of the impact (Duflo et al.,2006). The partial derivative refers to the impact of the programme or resource holding all other factors constant. The total derivative refers to the impact of the programme after that programme has affected all outcomes including other intermediate factors along the theory of change. For example, providing textbooks to schools could have the effect that schools reallocate their own budgets away from learning support materials. In this case, the partial derivative may be positive (the textbooks have a positive effect on learning) but the total derivative could be zero (if textbooks end up replacing other learning support materials). Similarly in the NSNP TOC, the provision of a meal may be positive however if this leads to parents not providing breakfast o learners the overall nutritional status of learners may not change and the intervention may substitute a previous allocation of resources.

An RCT measures the total derivative. However, if data on intermediate outcomes is collected the partial derivative can also be estimated. Both of these effects or impacts have implications for government planners and policy makers. The total derivative provides information on the 'real' policy impact, informing us about how the system and participants within the system would reorganise and respond or re-optimise when an input or resources, such as textbooks, are exogenously provided. As in our example, this would assist policy makers in realising that, although textbooks are a good resource in and of themselves, the change or lack of change by teachers, parents or learners when they have textbooks provided by the state limit or enhance the effect of textbooks in improving learning. Similarly although the provision of meals is important in alleviating hunger and improving nutrition, overall improved nutrition amongst learners requires continues provision of meals by households.

In addition to understanding the partial and total derivative when evaluating a TOC, determining the most strategic pathway in a TOC or attempting to understand why a policy or programme has not achieved the desired outcomes based on a TOC may be assisted by the use of the binding constraints approach when evaluating during or after implementation. The binding constraints approach has been developed by Haussmann et al (2006) and was also used by the so-called Harvard group that investigated South Africa's economic growth prospects. The approach essentially starts from the perspectives that although there may be many constraints, not all of them are equally binding. Applying this approach to education, if in a particular country schools were dysfunctional, or teachers were not teaching many of their classes, improving teacher subject knowledge may not bring much reward, even if this may indeed be a constraint. In the NSNP TOC example above, the outcome of improved educational performance may be hindered if the binding constraint is not hunger but poor teacher content of pedagogy. Thus although the lack of nutrition is a constraint to learning, the binding constraint would be outside the designed TOC and thus the desired outcomes would not be achieved. The binding constraints approach helps us to assess priority as the most binding constraints should be dealt with first.

Policymakers often have to choose which policy interventions or areas to focus on but due to a lack of rigorous information, the methods that tend to be used for prioritising, tend to be unsystematic and consequently inefficient. It is plausible to imagine that the casual impact between feeding and improved learner performance is difficult to measure although there is a body of knowledge from the health sciences linking nutrition, mental development and performance. In a particular setting, education outcomes may be constrained simultaneously by a variety of factors such as home background, school management quality, teacher effort, and teacher skill. In South Africa, as in many developing countries, there is evidence of all of these factors being present. However, simply observing these factors does not necessarily inform a policy maker as to which constraint to address first.

Prospective impact evaluations have the advantage of uncovering knowledge of the binding constraints within the system of interest, such as the education sector, where there are often strong relationships between multiple factors constituting system failures. For example, descriptive research shows low levels of accountability (e.g. NEEDU Report, 2013) as well as weak teacher capacity (e.g. Venkatakrishnan and Spaull, 2014) within the South African school system. Here, an intervention premised on an accountability TOC (such as rewards for improved learner performance) may not have any effect if the major binding constraint is a lack of subject knowledge amongst teachers. Holding teachers accountable may not change anything if teachers do not know what to do differently in the classroom. Therefore, by implementing new interventions, each with slightly different change theories, an RCT design ("playing with the levers") can help one figure out what the binding constraints are and in this way derive broadly instructive findings that will inform the general direction of intervention design in that setting.

The use of multiple treatment arms in an RCT, otherwise known as a factorial or cross-cutting design, allows for the evaluation of competing interventions in addressing the same policy problem (Svensson, and Pettersson-Lidbom 2008). This is important, not only because this allows for the identification of the binding constraints, but it provides the opportunity to assess the cost-effectiveness of alternative interventions. The most cost-effective ways to bring improvement are not always to tackle the biggest barrier; comparisons that identify the most cost-effective way of addressing a specific policy objective are therefore particularly useful for policymakers who work within limited budgets, as cited in the Department of Basic Education evaluation workshop report (2012).

Some critics of RCTs argue that most education interventions or programmes are too complex to evaluate through impact evaluations (Mouton, 2009). Designing RCTs to include multiple treatment arms and to collect information about intermediate outcomes along a specified TOC can go a long way towards addressing these criticisms.

2.8 The Technical Requirements for Randomised Control Trials

In the section above, the problem of selection bias and overcoming this through internally valid quasi-experimental and experimental methods was discussed. The benefits of specifying a clear TOC for understanding the binding constraints within a particular education policy setting have also been elaborated on. The next section discusses the statistical requirements for conducting RCTs. These are minimum statistical parameters that make it possible to use RCTs to measure impact.

2.8.1. Sampling: Sample Size, Statistical Significance and Confidence Intervals

In implementing interventions for research purposes it is not necessary to evaluate an entire population; information from a randomly selected, representative sample may be used to estimate population parameters. “Sample size as well as other key design features affect the power of experiments and whether one is able to reject the hypothesis of zero effect” (Duflo et al., 2006). A first principle of sampling is that the larger the sample the more precisely one will be able to generalise about findings to a larger population.

Understanding sampling as a statistical technique is critical for RCTs. Incorrect sampling results in an inability to infer the findings from the sample data to the larger population of interest. Key considerations for sample size are the level of precision required in the evaluation and the level of disaggregation that is expected when interpreting the results. In the South African schooling system, it is generally optimal to be able to disaggregate findings by province rather than only at the national level.

The context of the sector and country are among the variables that need to be considered when sampling. As a rule, the greater the variation in the population of interest the larger the required sample size. Interventions are generally targeted at specific sub-sets of the population and this directly affects sample selection. This sub-group may be classified according to socio-economic status, educational attainment, race or gender, amongst a range of characteristics of interest, and understanding these sub-populations and groups is important.

In South Africa, there is a broad range in learner performance levels across the system but differences between schools is larger than within schools. The intra-class correlation coefficient (ICC) is a statistical measure to reflect the proportion of overall variation that is between-school variation relative to within-school variation. An ICC of 0.6, for example, means that 60% of the variation in test scores is attributable to differences between schools and 40% is attributable to differences between learners within schools. Therefore, when a sample of learners is clustered in similar groups, such as schools, the sample size has two dimensions, namely the number of schools and the number of learners to be sampled within each school. Increasing the number of schools in the sample will offer more statistical power than increasing the number of individuals sampled within each school. As Duflo et al. (2006, p. 34) explain, "When group outcomes are correlated, data from another individual in an existing group provides less information than data from the first individual in a new cluster".

A further measure that may be used to improve the precision of estimates is stratification of the sample prior to randomisation. "Randomisation ensures that treatment and control groups will be similar in expectation but stratification is used to ensure that along important observables, this is also true in practice in the sample" (Duflo et al., 2006:36). Stratification is based on creating groups (called strata) within the sampling frame based on shared or observed characteristics and then selecting a specified number within each strata. In an RCT this means that a specified number of individuals or schools within each strata are randomly allocated to treatment and control groups. This improves the chances of treatment and control groups being well balanced. The strata may be grouped by SES, gender, language of learning and teaching or baseline test results. An additional reason for stratification is an interest in establishing effectiveness, or effects based on sub-groups of the population, in addition to the aggregated effect. Considerations when stratifying should therefore include sample size and disaggregation by strata so as to ensure sufficient statistical power to make inferences representative of the larger population within the strata at the conclusion of the experiment. In the South African public schooling system, schools are categorised by poverty quintiles, although these are not equal in size as the definition implies, since a larger proportion of schools are concentrated in Quintiles 1 to 3. When selecting a sample in this

context it is recommended that stratification should be undertaken along quintiles, especially considering the difference in size, SES as well as learner performance across quintiles.

Statistical significance, otherwise known as the alpha parameter, is a further critical component affected by sample size, effect size and power. Statistical significance is the level of assurance or precision of the measured effect. This alpha parameter refers to the need to reduce the risk of concluding that there was an impact when in truth there was no impact, i.e. that the effect size is significantly different from zero. In legal language, it may be understood as the risk of finding somebody guilty when in fact they are not guilty. Statistical significance is closely linked to confidence intervals, which set upper and lower bounds for the true effect value of an intervention. An acceptable convention is the 95% confidence interval – this is the range in which one can be 95% sure that the true population value lies (Taylor, 2014). Confidence intervals that are broad may tell us that the true value of the effect size lies between too broad a range to be meaningful. The acceptable width of a confidence interval is dependent on the effect size: if it is closer to zero, the confidence interval has to be narrower (Hopkins, 2000). Larger sample sizes generally narrow confidence intervals but as sample sizes are affected by financial constraints increasing the size of the control group rather than the intervention group may be more cost-effective, as the increased costs would be for data collection only, and not for the proposed intervention.

A further component to consider when determining a sample is statistical power. Statistical power is based on establishing how well the design of a study is able to distinguish the real effect of the intervention from a chance difference (Schochet, 2005:2). There is a, a close relationship between sample size and power; when the sample size decreases, the statistical significance levels decrease and similarly, the statistical power. Statistical power calculations are based on establishing a minimum detectable effect size (MDE) - this is the smallest effect one wishes to identify with statistical confidence. It is recommended that the minimum effect size should be the smallest effect size that is large enough to be cost-effective as a policy intervention. For example, a policy maker may decide that a treatment effect of 3 percentage points is too small to justify the cost of an intervention, but that an effect of 4 percentage points would be just large enough to justify the intervention. In this

case, the sample should be large enough to be sure to identify an effect of at least 4 percentage points but too small to be sure to identify any smaller effect should it be there. There is a trade-off, however, between cost and power: increased power requires larger samples but larger samples are more costly financially. Paradoxically, the notion of a minimum detectable effect size means that more expensive interventions might be less costly to evaluate, since a smaller sample will be required.

Power calculations are based on mathematical equations relating various relevant parameters, such as the MDE, intra-class correlation, alpha, power, cluster size, etc. However, some of these parameters may require a fair amount of guessing supported by prior experience. This includes having an idea of the mean and variance of the outcome prior to the experiment using prior data, preferably from the same region. When this is not possible in designing an intervention, one may specify the desired effect size using standard deviations. Statistical programmes would then use this information to calculate the sample size and other specifications. Power estimates that are considered to be adequate are generally between 80% to 90% although budget constraints influence the feasibility of this (Duflo et al.,2006). In practice, many evaluations have been conducted when they were underpowered, i.e. the sample was always going to be too small to provide statistically significant results. Unless qualitative methods are to be applied, this situation should be avoided.

The question of what determines a desired effect size is the subject of much debate in education, with no easy recommendations available. There is some consensus, however, that a rough estimate of one year of learning is equivalent to 0.4 to 0.5 standard deviations of test scores (Taylor, 2015). Determining the magnitude of an effect size requires a range of considerations including knowledge of the normal gains on a yearly basis through standard schooling practices as measured by learner assessments; overall performance trends based on school type, quintile and general profile; and information on overall systemic averages. The resources, time invested, and the cost-effectiveness of an intervention should also be part of the consideration.

Despite the fact that standard deviations provide a useful measuring tool, they are not without their perils. The main question facing the academic community is 'How standard is a standard deviation?' as standard deviations are dependent on the population being tested as well as the test instrument. They reflect a measure of dispersion based on the nature of the population participating in a test and thus the standard deviation in a test taken by learners in rural schools would be different from that of learners in urban schools or a combination of both using the same test. As such, simply comparing effect sizes in terms of standard deviations may artificially inflate the effectiveness of an intervention depending on how homogeneous the group is (McKenzie, 2015). This is particularly important when comparing the effect of interventions in a pilot or small sample against large-scale implementation where heterogeneous participants are more likely to be found.

The second limitation is based on test instruments, namely, that differently designed tests provide differently shaped distributions and resultantly, different standard deviations. This is further affected by the method, whether weighting or Item Response Theory (IRT), that is used to aggregate the test scores. Proposals to improve the interpretation of results include reporting multiple measures including impact size, standard deviations for overall tests as well as subsections. In addition, an important recommendation for the education sector for both policy makers and academics include adopting the standardised testing practice of repeating the same test items across various studies for effective comparison in interventions and research studies; and providing comprehensive information on how test scores were derived when reporting standard deviations or aggregate scores (Singh, 2015). The measurement challenges brought by a range of instruments and measures is an area that has recently been receiving attention, much can be said on this.

An alternative method of understanding effect sizes is comparing the findings of a variety of credible experiments in the same area to make sense of whether the impact is meaningful as proposed by Schochet (2005). This may be further assisted by the use of the same test items across various interventions as best practice that should be pursued by research bodies as described above. A more short-term approach may be item mapping, where individual test items are grouped against various skill and knowledge categories and then compared. This method which is used by National Assessment of Educational Progress

(NAEP) is the largest nationally representative continuing assessment in the USA provides a useful framework for comparison (NAEP) 2015). This method provides an opportunity to group and analyse questions targeted at the same skills or knowledge and thus improve the interpretation of the test results across tests based on the same skill sets and cognitive demand.

Kremer, Brannen and Glennerster (2013) present a cost-effectiveness analysis of 30 RCTs at primary school level based on reviewing different evaluations and their effect size compared to the cost. Their analysis compared gains in test scores per $100 and thus provides a benchmark against which effect size can be measured. Further discussion on costs I provided in this report.

A further alternative is assessing the long-term benefit of gains, such as possible future earnings. Several studies (cited in Schochet, 2005) suggest that an increase of one standard deviation in learner performance in mathematics or literacy results in an increase of 8% in future earnings. This kind of long-term effect could be used to gauge if the effect size of an experiment, beyond the immediate outcomes, is meaningful. A practical example of interpreting effect size was completed on the Mind the Gap study guide RCT (Department of Basic Education, 2013). A simulation was conducted to assess the proportion of learners that would have passed the National Senior Certificate (NSC) examination if they had received the Geography and Life Sciences study guides, based on a measured effect size of 2 percentage points gained by learners that had received these. Assuming all learners who had enrolled for these two subjects had scored an additional 2 percentage points, it was estimated that 5609 learners would have passed the NSC examinations if they had had the materials. This simulation was useful in conveying the significance of the study guides for policy makers. Although the effect size was 2 percentage points, in this context that measure is substantial.

Attrition rates should be considered as a further key factor in determining sample size. Power calculations should therefore be based on the expected realised sample rather than the initial sample size prior to attrition. If attrition is systematically correlated with treatment status this can introduce bias in the estimated treatment effect. The most

common example is where attrition takes place mostly in the control group, and the effect size may therefore be overestimated (Duflo et al., 2006). The reasons for attrition are worth considering, and building in methods to track attritors even after they leave experiments should be pursued where possible. Reasons for attrition may include factors like mobility rates across the country or region, learner drop-out or churning in the schooling sector, as well as the intervention administration or data collection point. If data collection for an intervention takes place at a central location and a significant level of attrition might be expected, developing the scope and budget allocation to enable the data collection of attritors in their homes may be a useful investment to maximise the statistical power. This may also be used to establish the rationale behind opting-out. Tracking attritors requires prior planning as well as an adequate data collection exercise of background information to allow for tracking. An alternative to tracking all attritors where the cost is exorbitant, is tracking a random sample of attritors for in-depth follow-up. The weighting of this second sample should be adjusted in the final calculations, and best practice is to report the attrition rate and measures taken to address this in detail in the final report. In the absence of follow-up options, statistical methods which factor in attrition may be used (Duflo et al., 2006).

Determining the analysis method through the use of either Intent to Treat (ITT) or Treatment of the Treated (TOT) influence the sample and interpretation of results. Partial compliance, measured through fidelity in the enactment of the intervention components, such as the use of lesson plans, working through reading material or using the assessment and feedback components of an intervention in the schooling context, may have an effect similar to attrition when measuring impact. Historical information on similar interventions and a good understanding of the practice and constraints in the education system would be necessary in order to plan and account for the expected levels of fidelity and adjust the sample size accordingly. The analysis methods may include using ITT estimates where everyone assigned to receive treatment would be included in the analysis, potentially underestimating the size of the impact or TOT when the estimated treatment are limited to those treated (Vivalt,2013). With either method, the effect on the sample, attrition and effect size are serious considerations.

2.8.2. Baseline Data Collection

Unless routinely collected administrative data of sufficient quality and relevance is available, the collection of data is integral to the RCT methodology, as it is through the newly collected and generated data that causal impact is measured. The question of whether to collect baseline data is based on several factors that should be carefully considered, although strictly speaking, it is not necessary since randomisation means there is no reason to expect any differences between treatment and control groups apart from the impact to the intervention. The advantage of collecting baseline data, is that it facilitates the collection of variables that will reduce the variability in interpreting the endline data and this allows for a reduced sample size and provides assurance that randomisation was done correctly, as observable characteristics may be compared between the treatment and control groups. Furthermore, the opportunity to factor in the prior knowledge of participants when analysing the treatment effect provides an improved measure of the effect size of the intervention. It also allows for the examination of the interaction between prior knowledge and the intervention; it allows for more complex analysis of data which may indicate that the intervention works best for a specific subset. Such information may, however, be available in existing administrative data.

Baseline data also contributes to external validity. If the selected sample is similar in characteristics to the larger population beyond the experiment, it may be possible to argue that the findings and impact from the intervention are valid for other population groups beyond the experiment. Furthermore, when conducting research in a relatively new research area where little is known about the status quo, baseline data-collection allows for a further contribution to the body of knowledge, providing information where none was previously available. For instance, the Early Grade Research Study (EGRS) RCT targeted at Grade 1 in a sample of schools in the North West province, mentioned earlier, is a case in point. In the South African schooling system, the knowledge and skills level of learners at the start of Grade 1 are not measured nationally and there are no standardised national assessments at the end of pre-schooling. The baseline data collection through the EGRS will therefore not only contribute to the analysis of the RCTs results, but provide new

information on the levels of knowledge and skills of 5 to 6 year olds at the beginning of formal schooling in South Africa.

McKenzie (2012) argues that the administration of a baseline assessment followed by a once-off endline survey are appropriate for highly auto-correlated and relatively precisely measured outcomes, as is usually the case in education settings when using test score data. However, in many clinical trials or other settings where auto-correlation is low, such as in RCTs on spending and financial practices, multiple measurements at short intervals allows for better data to average out noise and increase power. In light of endline data collection, the baseline data collection exercise affords the experiment a practical learning and preparation opportunity. Lessons learnt may be critical in ensuring that valid data is collected at the final data collection exercise.

The use of administrative data as an alternative to baseline data collection is often a cost-effective method of collecting information on the treatment and control groups but the reliability, completeness and quality assurance practices in place when collecting such data should be fully understood prior to the use of this data. In addition to its possible use as baseline information, administrative data may also be useful when analysing the findings from the research: merging the RCT data with administrative data may provide additional information as well as provide an opportunity to complete further analysis that would not be possible without using the combined dataset.

2.8.3 Data analysis

The specific techniques and statistical formulas that are used to analyse data and interpret results from RCTs are beyond the scope of this study. One critical aspect worth discussing however, is the manner of determining which factors to control for (already mentioned above with reference to the contribution of these variables to variability). Good practice is to report the key findings from experiments with and without (raw) controlling for several

variables in order to reduce bias in the results and data-mining [3] (Duflo et al.,2006). In clinical trials, the standard practice is to establish ahead of time protocols that indicate how the data will be analysed. This is known as a pre-analysis plan. This plan provides information on the specific research questions as well as how the data will be analysed in order to reduce the possibility of biased analysis and reporting. Similar practices are increasingly being adopted in the social science RCTs to improve the integrity of the results published. Failure to follow treatment protocol or partial compliance by participants is a further factor to consider when analysing RCT data, as mentioned above. Those selected for treatment may not always comply with the instruction for the intervention, thus introducing a deviation from full randomisation and an element of self-selection. The availability of data on compliance may be used in statistical practices to deal with this. (Svensson and Pettersson-Lidbom, 2008).

2.9 Overcoming Publication Bias

Publication bias refers to the tendency amongst authors and publishers to favour certain results from research, especially large positive effect sizes. This leads to an overly favourable review of the benefits of categories of interventions across the literature, creating an unrealistic expectation of drastic changes from interventions. The publication of a pre-analysis plan can go a long way toward reducing this practice. Secondly, the methodological rigour and required effort of RCTs as well as their very nature as experiments, often warrant publication or sectoral discussion even when a null result is obtained. RCTs are often completed in a collaborative manner with funders included requiring reports documenting the findings which in turn form part of knowledge generated and being referred to in future work. In addition, when randomisation is done properly the reported results from an RCT are plausible; even when results are unexpected, measurement error is not likely to be the reason and thus publishing the results is more likely to be done by both the principal investigators and journals. This is in contrast to less rigorous research methods – a data analyst looking for interesting correlations within a cross-sectional dataset with numerous

[3] Data mining is the manipulation of data on many different characteristics in order to find statistically significant relationships. It entails cherry-picking of relationships to falsely promote a positive or negative effect without reflecting on the overall findings (Glennister and Takarasha, 2013)

variables is unlikely to go out on a limb and write a paper about textbooks not having an impact on learning outcomes than she might be if she found a large positive correlation. Glewwe, Kremer and Moulin (2007), however, published a famous paper using an RCT showing a zero impact of textbooks in Kenya. The literature based on RCTs may not be completely free of publication bias but the use of RCTs goes a long way towards reducing the probability and prevalence.

2.10 Critiques and Caveats of RCTs

The critiques and caveats of RCTs may be categorised under three broad themes, with some overlap. To some extent, this chapter has engaged with the first of these through the arguments for the use of RCTs in view of the limitations of other available methodologies. A second criticisms levelled can be categorised as epistemological, contention on the basis of particular theories of knowledge arguing that RCTs are essentially an invalid methodology for the social sciences. The international and national debate followed by three main points of this critique, the "gold standard" RCT rhetoric, the "blinding" criteria, and the ethics of RCTs will be discussed below. This will be followed by an exploration of the third set of criticism, which consist of various technical methodological and utility caveats

2.10.1 Epistemological Criticism

The abuse of the use of RCTs by various agencies that may benefit financially or academically has contributed to an epistemological aversion to RCTs; and the fact that these include international agencies such as the World Bank and the USA Department of Education's Institute of Education Sciences (Deaton: 2009; AEA: 2003) makes this point that much more serious. A case in point was a 2004 debate on how to practice evaluations in Claremont, USA where the full $500 million budget from the USA Institute of Education Sciences was awarded to a select number of researchers who are proponents of RCTs (Donaldson and Christie, 2004). This award highlighted a funding bias in evaluations resulting from a decision that RCTs are the only acceptable research method. This standard was applied regardless of the nature of the research question, its purpose and current or existing knowledge available. It is clear in this case that there may be a conflict of interest or

unnecessary pressure emerging from assertions made by the USA government, proponents of RCTs and the outcomes of RCTS. The unnecessary pressure of utilising a single method considering the resources invested and promises of establishing causality are not ideal for unbiased scientific discovery (Deaton, 2009). However, to end the debate at this juncture would unnecessarily restrict the use of a range of technically viable methods such as RCTs and constitute an unwarranted limitation in addressing social policy problems.

The debate in South Africa on RCTs or other experimental and quasi-experimental methods within the education sector particularly amongst education academics is very limited. The critique by Mouton (2009) argues that impact evaluations are more geared to measure simple, linear relationships rather than complex social programmes such those in education, which are implemented by various agents and may have multiple simultaneous causal strands. A further distinguishing factor is non-linear causality and emergent priorities- defined as the emergence of specific outcomes and the means to achieve them during the intervention rather than prior. Bearing this in mind, there is an additional dimension of gains or impact accumulating or diminishing over time thus complicating the evaluation process. An alternative approach proposed by Mouton (2009) is the "evaluative case study" design that allows for weak causal claims based on persuasion. At the heart of this critique is a belief that the social sciences cannot simply appropriate the research methods developed in other disciplines such as medical and economic fields, such as in the case of RCTs. The critique warrants careful consideration when implementing quantitative methods to evaluate social problems, but the blanket statement that such methods are wholly inappropriate seems extreme. Neither the critique or proposed alternative methodology - evaluative case studies – provide a compelling rigorous argument in terms of the key contributions of RCTs, namely addressing selection bias and providing a valid estimate of the counterfactual using large sample sizes that represent the population of interest. The proposed method lacks both internal and external validity and therefore does not respond to the real questions of policy makers. Although the complexity of the educational context is a key consideration, policy makers require evidence from large scale evaluations of interventions with confidence that the estimates of impact are relevant beyond a single case study or schooling context.

RCTs are sometimes referred to as the “gold standard”, a label that originally referred to randomised double-blind placebo control (RDBPC) studies in epidemiology (Misra, 2012; Clay 2010), which enabled researchers to determine whether exposure to a specific treatment was directly responsible for an outcome. This was informed by a causal hierarchal arrangement of studies based on the evidence they can provide. Unfortunately, the phase “gold standard” connotes perfection, which is unhelpful since RCTs, similar to all other methods, are not a perfect methodology, especially outside clinical trials where the RDBPC criteria is not met. It is fair to say that the “gold standard” label has been applied inappropriately to advocate the use of the RCT methodology beyond its capabilities. A more accurate description would therefore be that properly implemented RCTs provide unbiased estimates of impact consistently (Deaton, 2009).

It is acknowledged that the RDBPC criteria, although ideal, are not always possible or ethical even in medicine, and therefore other alternatives or adjustments are made. These include distinguishing between a placebo effect - a psychosomatic effect brought about by relief of fear, anxiety or stress because of receiving a treatment- and no treatment. In order to determine the treatment effect, a placebo must be given, and the effect of the trialed treatment must be better than the placebo. An alternative to the placebo is the positive active control utilised where the use of a placebo is unethical. An example is the case of treating the Human Immunodeficiency Virus (HIV): because of the seriousness of the illness it is not acceptable for participants in trials to not receive real treatment. Participants are therefore given the standard treatment that is accepted as effective, otherwise referred to as a positive active control. HIV treatment trials therefore compare the standard treatment to the new proposed treatment (Misra,2012).

Similarly, in the case of education research, RCTs generally evaluate an alternative to a positive active control- if proposed interventions produce better results than what is currently available and offered in the standard schooling experience. This framing of RCTs addresses two fundamental issues: firstly, there is a clear impetus to measure the standard practice, which implies that there is an acceptance that standards can and should improve; and secondly, the comparison against a positive active control clearly communicates that changes and ‘’improvements’’ within the field should be based on rigorous evaluations. It

cannot be assumed that all new proposed treatments are better than the standard practice. These are important lessons and practices that should be adopted in the field by researchers and policymakers.

A further component of epidemiological studies, specifically in light of the RDBPC criteria, is the requirement for participants to be blind to the treatment. Misra (2012) points out that when the outcome of trials can conceivably be affected by the patient's or investigator's expectations, then "blinding" is important. There are three types of "blinding", namely; single blinding - when the patient is blind; double blinding -when the patient and the investigator or data collectors are blind; and triple blinding: when the patient, investigator and data collectors are blind. An important instance in medical RCTs where the triple "blinding" criteria was not applied was in trialing the impact of male circumcision on the rate of new HIV infections in South Africa, Kenya and Uganda between 2002 and 2007. Participants could not be blind as the treatment - circumcision - entailed a surgery. The treatment group was circumcised at the start of the RCT, while circumcision was delayed until the end of the trial for the control group. The RCTs in all three cases were concluded early as intermediate findings showed a substantive reduction in the rate of new HIV infections for the treatment group (Harrous-Paicheler, undated; ClearingHouse, undated). Although the results and policy recommendations are contentious, these cases clearly show the possibility of a reasonable waiver of the "blinding criteria". Nor are these cases isolated. A peer reviewed systematic review of 68 HIV/AIDS intervention or prevention RCTs conducted between 2004 and 2008 in Africa found that only 18 of these studies fulfilled the triple-blind criteria, 7 fulfilled the double-blind criteria, and the remaining studies either fulfilled the single-blind criteria or none (Zani, Pienaar, Oliver and Siegfried, 2011). It is worth noting at this juncture that 66 of these 68 RCTs secured national ethical clearance, while 57 were granted non-African ethical clearance (where the non-African countries involved included the United States of America (USA), Switzerland, Canada, France and Denmark) in addition to national ethical clearance. The argument that RCTs in education settings do not have the same rigour as RCTs in the medical sciences is perhaps not as clear as one might expect prior to finding out a bit more about how they are actually conducted in the medical field.

The criteria for the requirement to be “blind” is clear in the medical literature and refers to instance where information on the trial would affect expectations in a way that would influence the treatment effects. However, these premises do not always apply in education. For instance, if the outcome measure in an RCT is reading, even if learners are aware of the expectation that an intervention should assist them in reading, the expectation to read would not translate directly into reading because of the complex skills and knowledge required to master reading. Outcome measures in education are often difficult to master; the failure of the system to convert current inputs and resources into meaningful learning outcomes is testament to this. It follows that the argument that without strict adherence to “blinding”, RCTs in education are not valid is clearly incorrect. Nevertheless, care should be taken and efforts made for data collectors to be “blind” where possible, in order to eliminate any bias in their assessment and data collection process.

The issue of ethics is often raised when discussing RCTs. The two main concerns are when participants in RCTs are asked to accept a burden or risk for the purposes of research which will benefit others, and when the control group might feel that they are being deprived. There are several points that can be made in defence of the ethics around RCTs. The primary reason for conducting research, and RCTs specifically, is to improve the condition of society by establishing if proposed interventions or treatments are effective of if they are more efficient than what already exits. It could even be argued that it is unethical to implement programmes or interventions that have not been scrutinized and evaluated in a rigorous manner. Indeed, the question of ethics is not limited to RCTs but applies to all research and practice, particularly where individuals or groups participate or provide information. The balanced view should therefore be that there are ethical and unethical approaches that may be followed in every research methodology.

Some have argued that RCTs are a justified practice conditional on informed consent and equipoise - a state of regarding two treatments as equal in prospect (Edwards et al., 1998; Grady 2008; Mfutso-Bengo, 2009). It is these two conditions that are highlighted in the academic literature, with the main concern being informed consent. There is consensus that participants are unlikely to fully understand all the information provided; hence, fully informed consent from all participants is unlikely to be the case even when consent forms

have been signed. Edwards et al. (1998) propose three possible responses to this problem: declaring all RCTs to be unethical unless participants are themselves sector experts; abandoning the requirement for informed consent and instead relying on ethical committees to provide consent for all participants; or retaining the spirit of informed consent while taking practical measures to maximize the understanding of participants. The third option seems to be the most ethical and practical response as it acknowledges the limitations of the current practice of informed consent, and proposes efforts to enhance these while facilitating the scientific purpose of research.

In the education sector, the approach to informed consent is usually a hybrid between the second and third option above. As custodians of the education system, the DBE and provincial education departments (PEDs) are mandated to develop policy, govern and monitor the delivery of high quality education. In fulfilling this mandate the DBE is required to plan, evaluate and report on the medium to long range performance of the education system, including comparative analyses and research co-ordination in support of overall sectoral goals. The practical implications of this include conducting research using a variety of credible methods in schools without attaining direct parental consent as parents enter into the education compact by enrolling their children in public schools. In an effort to more fully realise the spirit of informed consent, however, the practice is to inform the various structures within the sector including districts and school managers. Furthermore, researchers should maximise opportunities and efforts to provide parents with meaningful information pertaining to specific research and evaluation efforts. This may be done by sending consent forms to parents for signature, as has been done in the Early Grade Reading Study RCT. In the two RCTs reported on in this study, ethical clearance was also obtained through the IRBs of Research Institutions involved in the project.

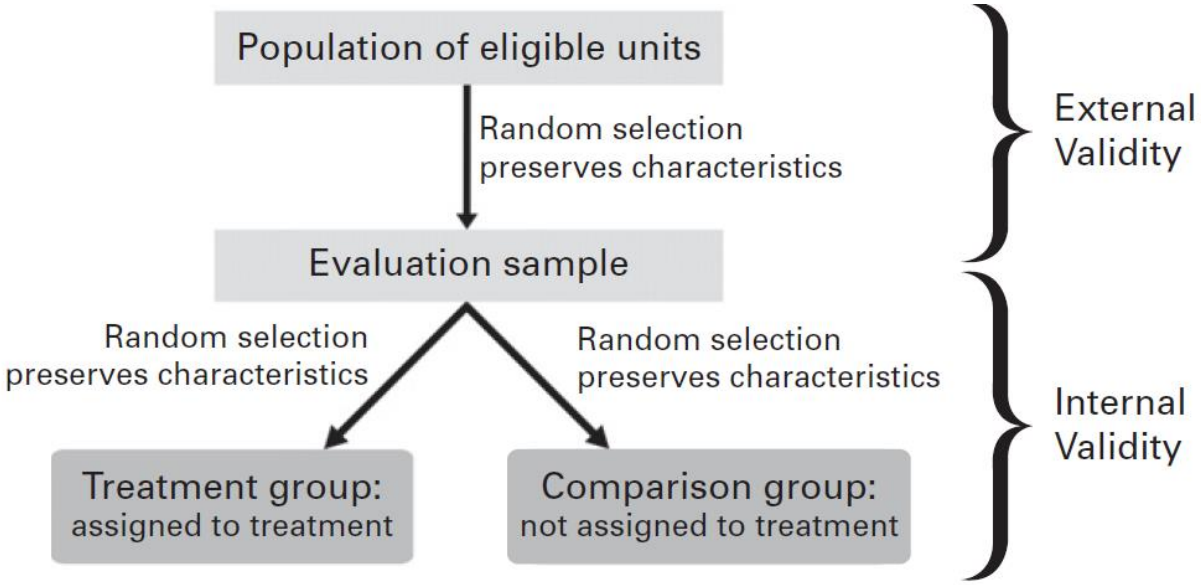
As discussed above, the epistemological debate around RCTs appears to be rather superficial without any engagement with the substance of the methodology. The abuse of RCTs by various agencies that may benefit financially or academically has contributed to this epistemological aversion; and the fact that these include international agencies such as the World Bank and the USA Department of Education's Institute of Education Sciences (Deaton: 2009; AEA: 2003) makes this point that much more serious. A 2004 debate in Claremont on

how to practice evaluations highlighted one of the aversions to RCTs as the only acceptable research method is the allocation of the full $500 million budget from the USA Institute of Education Sciences to a select number of researchers who are proponents of RCTs (Donaldson and Christie, 2004). There may be a conflict of interest or unnecessary pressure emerging from assertions made by the USA government, proponents of RCTs and the outcomes of RCTS. The unnecessary pressure of perfection considering the resources invested and promises of establishing causality are not ideal for unbiased scientific discovery (Deaton, 2009). However, to end the debate at this juncture would unnecessarily restrict the use of a range of technically viable methods such as RCTs and constitute an unwarranted limitation in addressing social policy problems.

2.10.2 Caveats of Randomised Control Trials

The usefulness of RCTs has been clearly demonstrated in this chapter with the main benefits being statistically significant estimates of causal impact that are based on large samples and therefore can be extrapolated to a larger population. Nevertheless, there are several limitations that should be understood and critically engaged with in order to guard against inappropriate RCT use, based on unfounded expectations that they are either silver bullets in education research or bear a unique stamp of approval prior to the implementation of any intervention. The challenge of ensuring internal validity, that is, the need to create a valid estimate of the counterfactual, has already been discussed. Issues of external validity defined as “the ability of an experiment to replicate its results in other contexts” (The Takshashila Blog, 2015, p.1) are discussed in the section that follows. The discussion also focuses on concerns on how data is analysed and lastly, on the interpretation of results. The figure below demonstrates how various aspects of RCTs may be categorised into these two validity concerns.

Figure 2: Random sampling and randomised assignment of treatment



***Source**: Gertler et al.,2010)*

2.10.2.1 External Validity

The main critique is about the replication of results in different contexts, context does not only refer to geographical location, but also to time (Lemons et al.,2014). There is tension in balancing the reality of RCTs as controlled experiments carried out in a specific context, while maintaining that the experiment assumption and findings will hold in the larger population outside the controlled experimental environment.

The selection of a sample that is representative in a manner that is important to the system and is generalisable across a bigger population is the first effort that should be made to address the primary concern of external validity. The availability of administrative and assessment data and an understanding of the priority policy concerns in the education context would enable this. A practical step would be consultation with a range of stakeholders including policy makers at the various levels of government, academics across various institutions and fields, as well as experienced non-governmental organisations (NGOs). Even as international organisations or specialists are co-opted to conduct RCTs in developing countries, there is necessarily a requirement for strong local, meaningful engagement and a wealth of knowledge on the context and sector. If the sampling is done

well, and is sufficiently large, it should be plausible to argue that the findings from the sample are reflective of the larger population.

The second critique is about the effect of the implementing agency on the results: RCTs are often implemented by external professionals, including NGOs, academics and sector experts, but in scaling-up, the common practice in government is the cascade model which entails implementation by government officials removed by two or three levels from the RCT. These officials may not have the same skills or resources as the RCT implementers and may similarly not have the required appreciation of the intervention to maintain fidelity to the design and thus compromise the expected effect. The high levels of planning and control in conducting RCTs are often not translated into policy or practice (Deaton,2009).

Incorporating alternative models of scalability when designing RCT interventions into the post-experiment phase would largely address the second concern. This could be further enabled by developing partnerships with the different levels of government, from the initial implementation phase, to enable a transfer of skills and establish operational norms amongst the relevant officials who would be responsible for implementation. Furthermore, explicit capacity building activities should be considered upfront with a clear allocation of funds.

The third criticism is the experimental effect on findings, the controlled experimental conditions of RCTs may create the Hawthorne and John Henry effects. Both the intervention and comparison groups may alter their behaviour as a result of being observed and thus bias the findings. This is often the case when the experimental groups are in close proximity to each, or interventions are resource intensive. Although the observed changes may be misleading, the complexity in changing learner outcomes reduces the sporadic short-term reactions that would be measured. This implies that the tests and data collection instruments used in the RCT should be sophisticated and comprehensive enough to measure short-term and long-term impact on behavioural and learning outcomes. Communicating the value of measurement and experimental nature of the interventions may be useful in encouraging participants to behave normally. Lastly case studies and

qualitative research may assist in explaining the classroom-level changes measured in the impact of the RCT.

2. 10. 2.2 Data analysis

The second main area of concern is focused on the analysis of data and an overestimation of impact. The first concern in this area is misinterpreting the effect size based on the practice of calculating and reporting on the mean rather than the median. The average is affected by outliers that may skew the estimate of gains or the impact of an intervention when the estimate of the majority of participants may be substantially smaller than the average (Deaton, 2009).

In response to this concern, the first aspect would be ensuring that the sample is large enough to reduce the effect of outliers. The sample should also be large enough to study the differential impact of the same intervention on different sub-groups of the sample. This is known as Heterogeneity and refers to the diversity of impact due to different abilities, development or exposure. This nuance in the findings would assist policy makers in determining the appropriate subgroup in the population that would benefit the most from the interventions evaluated. It would also allow for an in-depth understanding of the interventions impact, beyond overall sample averages (Deaton, 2009). The collection of baseline data would also assist the accuracy in measurement as interaction with the baseline performance would provide estimates of relative gains made by individuals considering their starting point prior to the intervention.

Thirdly, it is good practice to investigate and report on numerous other more complex model specifications beyond the mean, such as tests for heterogeneous effects, effect on the median or other points in the distribution of outcomes, including other covariates for the sake of improved efficiency. Similarly, when compliance with treatment assignment is imperfect one may choose to use an Instrumental Variables approach to estimate the effect of the treatment on the treated, as opposed to the Intent to Treat estimate. A risk with these approaches, however, is of data mining – after enough alternative specifications there is a good chance that a statistically significant result will be found. This risk can be mitigated

through the use of a pre-analysis plan (published prior to data collection) which provides information on the specific research questions as well as how the data will be analysed in order to reduce the possibility of biased analysis and reporting. An explicit upfront effort to understand the fraction of those that made positive gains in an RCT through the analysis of heterogeneous effects may also be useful in presenting the nuances of implementing specific interventions and the specific beneficiaries rather than average gains. This does not negate the value of the RCTs but rather points out that RCTs are not a silver bullet and may be enhanced when a range of approaches or methods are used.

2.10.2.3 Interpreting the results of RCTs

The challenge of interpreting effect sizes in relation to standard deviations has already been discussed in this chapter. This discussion is therefore based on the critique for interpreting the results from an RCT. The primary concern in this area is overconfidence in RCT results. The main offenders are often policy makers and practitioners who do not engage with the detailed content and nuances and thus incorrectly extrapolate based on RCT findings. A second concern in interpreting the results is about intra-hypothesis validity defined as the validity within the hypothesis; recognising that there is no single version of the hypothesis or product but a plurality in how the same hypothesis could be evaluated. It is the recognition that a specific RCT is testing a single version of a Theory of Change using specific materials and implementers and as such, conclusions on the findings should be limited to that combination of resources or design (The Takshashila Blog, 2015). The third concern is that RCTs are able to measure impact but do not provide information on why things changed and this makes interpretation difficult as the immediate question policy makers need to respond to after understanding what interventions worked is why this happened.

Researchers should generally be careful in how they interpret the findings from RCTs but specifically refrain from making broad statements about the effect of teacher training for example, based on a single model of teacher training implemented in a specific RCT. The theoretical underpinning of all interventions should be explicit. It should be clear what Theory of Change is being tested. RCTs may, by design, respond to the question of why and not just what through combining theory, high quality assessment practices, and qualitative

components in the study to provide insights on why specific interventions worked while others did not.

Furthermore, in order to strengthen the interpretation of the findings of RCTs, internal replication should be introduced. The intention of replication is to provide a critique or support findings of the original study. If it is possible replication studies could evaluate the continued implementation of interventions that have been scaled-up from an RCT (Aiken and Davey, 2015). The International Initiative for Impact Evaluation (3ie) has recently introduced this approach as one of the pillars of their work. An example of this was a 2015 3ie grant for an independent researcher to replicate the findings of the highly influential Miguel and Kremer (2004) RCT on deworming in Kenya, given the data available. The main finding from the RCT was that deworming has a cumulative effect of 14 years on health and educational outcomes in terms of participation in schooling. This main finding was supported by the replication but it emerged that the estimate of the externalities amongst schools within 3-6kms of those treated was incorrectly reported. An error in writing up the analysis resulted in the estimates for the 12 closest schools which fell into a 3 km radius being reported as the effect on schools within the 3-6km radius (Aitken et al.,2014).

The main lesson to be learnt is that all research should be scrutinized regardless of its nature, RCTs may be a rigorous method but researchers implementing these are capable of making errors. Where possible, replication studies of the same or similar programmes in different contexts are encouraged to test generalisability and establish external validity (Glennerster and Takavarasha, 2013). Comparing RCT findings to those of quasi-experimental methods should also be employed to explore differences in estimates and establish patterns with large discrepancies leading to more in-depth analysis and exploration to account for these.

In conclusion the critique on RCTs emphasizes that researchers and policy makers should carefully consider the findings presented from RCTs and all other research methods and make every attempt to use a broader evidence-base than a single study or method as well as apply their minds when making policy decisions, considering the limited information they have and the limitations that are inherent in all evaluation methods.

CHAPTER 3: LITERATURE REVIEW

3.1 Introduction

Education research has predominantly used theoretical analysis and qualitative empirical methods, but the use of quantitative methods has been on the rise particularly over the past 15 years. There is widespread use of RCTs internationally, especially in the USA's education system (which since 2003 has been privileging experimental and quasi-experimental designs over other methods for purposes of evaluation funding competitions). This has been expanded by the establishment of international organisations such as the Abdul Latif Jameel Poverty Action Lab (J-Pal) and the International Initiative for Impact Evaluation (3ie). J-PAL was established in 2003 and has partnered with numerous NGOs, government institutions and private firms in both developing and developed countries, aimed at generating rigorous evidence about programme effectiveness that can inform policy reforms (J-Pal, 2012). The International Initiative for Impact Evaluation (3ie) has awarded over 200 grants in over 50 countries since it was founded in 2008. Their focus is funding impact evaluations and systematic reviews to generate evidence on the impact and effectiveness of development programmes (www.3ieimpact.org).

These organisations are a reflection of international agencies' promotion of quantitative methods, and of the resources invested and collaborations in the impact evaluation environment. As a result there is a large body of RCTs internationally in the broad field of development, with a substantial proportion of these in the field of education. Locally however, education research through experimental or quasi-experimental methods has been limited. With this in mind, this chapter provides a summarised review of some of the most significant RCTs in education from the international and local literature. There is also a specific emphasis on RCTs in reading. Insights from the literature are also discussed, and a critique of the studies is provided.

To narrow the parameters of the broader literature two main themes inform the literature discussion, namely; school participation and access, and learner performance. The discussions have been mainly drawn from five of the foremost systematic reviews on

experimental and quasi-experimental methods published in recent years, namely; Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments by McEwan (2015); the Challenge of Education and Learning in the Developing World by Kremer, Brannen and Glennerster (2013); What Education Policies and Programs Affect Learning and Time in School in Developing Countries? A Review of Academic and Aid Agency Evaluations from 1990 to 2014 for the Swedish Expert Group for Aid Studies by Damon, Glewwe, Wisniewski and Sun (2015); Interventions for Improving Learning Outcomes and Access to Education in Low- and middle-income Countries: A Systematic Review for 3ie by Snilstveit, Stevenson, Phillips, et al. (2015); and the J-Pal Evidence-Based Education Policy Making and Reform Conference (2012).

The second section in this chapter focuses specifically on four prominent international RCTs on reading; namely an RCT on textbook provisioning in Kenya by Glewwe, Kremer and Moulin (2007); an RCT evaluating the relative efficacy of a literacy skills development program in Mumbai by He, Linden and MacLeod (2009); an RCT on community participation to improve reading outcomes in India by Banerjee et al. (2010); and an RCT evaluating the impact of a one-month "read-a-thon" program in the Philippines by Abeberese, Kumler and Linden (2011).

This study attempts to respond to two research questions, specifically; presenting an argument for the value of RCTs as a methodological option for education research geared towards informing policy and illustrating this through selected case studies. As such one of the limitations of the literature review is the exclusion of a discussion on significant debates on the main theories on literacy, home language, and resources in reading. These discussions remain important in the broader education discussion, and this limitation is not a reflection of the relevance of these components in early grade reading literature.

3.2 International Literature on Experimental and Quasi-experimental Studies in Education

The largest body of literature is on school participation with a range of interventions geared at developing the most efficient mechanisms to improve access and participation, and reduce drop-out. These kinds of interventions are most effective where there is a low

demand for education, barriers to access, or high levels of poverty resulting in a trade-off between school attendance and earning income. Conditional grants, cash transfers, reduced fees and merit scholarships emerge as the monetary kinds of interventions that are effective; while improving child health and nutrition through providing meals, and reducing distances to schools have been identified as the most effective school-based interventions. Providing parents with information on the returns to schooling has also proven to be a successful intervention (Kremer, Brannen and Glennerster,2013; Damon et al.,2015).

The most renowned conditional grant programme, PROGRESSA, was implemented in Mexico in 1998. The programme offered conditional grants and medical care if learners attended school consistently. The impact of the programme was increased participation in schooling by 14.8 percentage points for females and 6.5 percentage points for males. Following PROGRESSA's success, more than 30 countries have implemented similar interventions which have been subjected to RCTs that have produced similar results (Kremer et al.,2013). In countries such as Kenya, where participation in schooling is differentiated by gender, similar interventions were found to be successful.

The studies focusing on school-based health and nutrition in order to improve access and participation include interventions such as deworming, malaria prevention and control, the provision of meals and incentives for anaemia reduction (Snilstveit and Stevenson, 2015; McEwan, 2015). There are hardly any programmes that combine interventions targeted at health and nutrition with those targeting curriculum support. McEwan (2015) points out that the gains to schooling are limited to an estimated average effect size of 0.04 which is statistically different from zero at the 90% confidence interval, when providing food, beverages, or micronutrients in schools. One of the most prominent studies in school-based health is the deworming RCT, discussed above, that was implemented in a densely populated farming district in Western Kenya between 1997 to 2001. The intervention was implemented in 75 primary schools with 30 000 learners. The impact of the intervention was a 31% reduction in the worm load of the treatment group, a reduction in absenteeism by more than a quarter for participants and a statistically significant benefit for those within 3-6 kilometres (kms) of the treatment schools.

Given the low unit-cost of deworming, the $100 invested resulted in a cumulative effect of 14 years of additional schooling (Miguel and Kremer, 2004) making deworming a cost-effective intervention. However, the systematic review by McEwan (2015) founds an average effect of deworming on learning outcomes of close to zero across 7 experiments. The interventions that have improved school access and participation have usually not made a substantial impact on learning outcomes, except in marginal gains made from improved participation as in the deworming RCT where there was previously very low attendance (McEwan, 2015). Access to education is a necessary condition but is not sufficient to result in improvements in economic attainment, individual earnings, distribution in income and economic growth. The level of skills and knowledge learnt through education are the fundamental measure of education. Thus the focus of recent interventions has been on improving learner achievement.

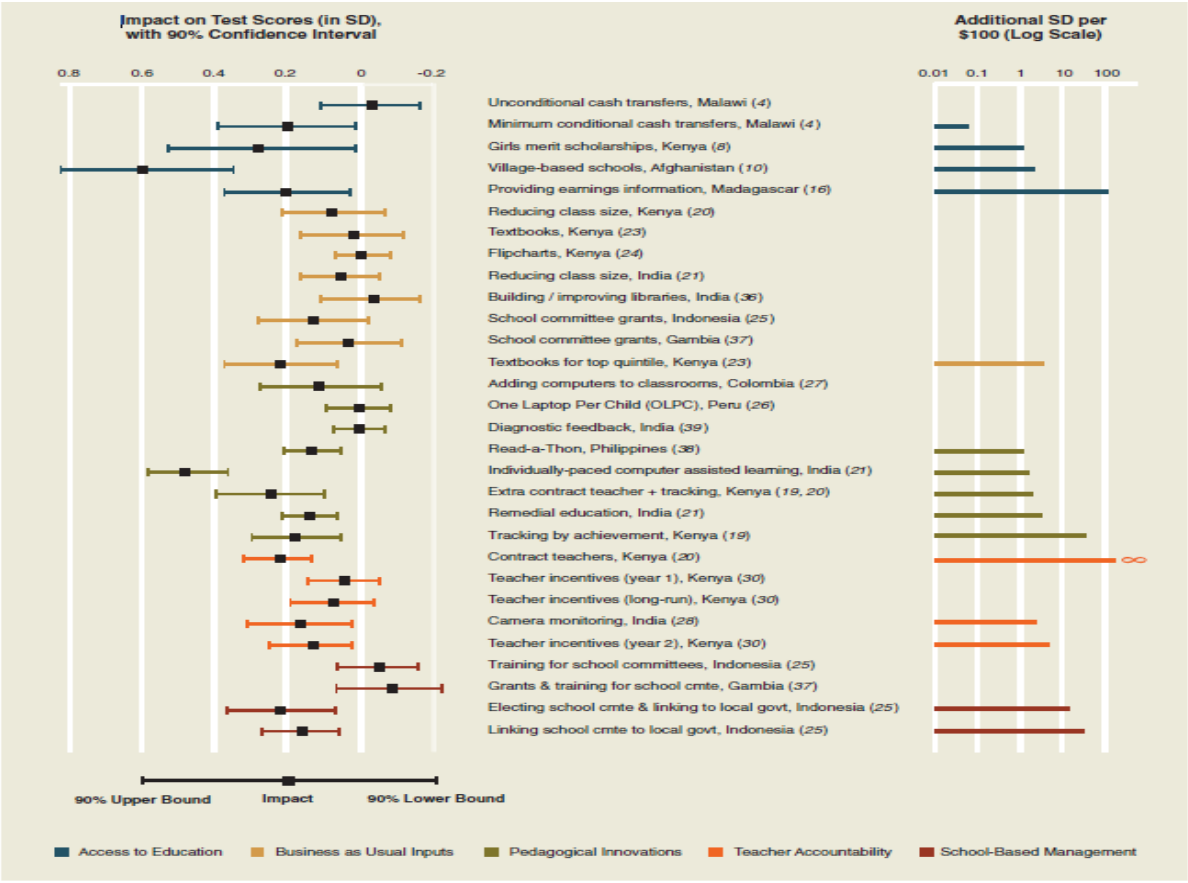
McEwan (2015) identified the largest average effect on learner performance as computers or instructional technology (0.15), teacher training (0.12), smaller classes with ability grouping (0.12), contract and volunteer teachers (0.10), student teachers or performance incentives (0.09) and incorporating instructional materials into teaching (0.08). Based on 18 studies, Snilstveit et al. (2015) point to structured pedagogical interventions as having the largest and most consistent positive average effects (0.23). These consist of new content focused on a particular topic, materials for students and teachers, and short term training courses for teachers in delivering the new content. The main critique to these interventions is the difficulty in isolating the specific effect of different components. McEwan (2015) highlights this difficulty when reducing class size by providing temporary teachers as both the effect of temporary teachers and reduced class size could be the cause of improvement. Furthermore, the contract-based employment status of the temporary teachers often creates a different dynamic to the permanent employment status of average teachers, which means that different incentive systems may be at play and therefore the positive effect may not be applicable to the broader system unless teachers are temporary hires.

A final resource that was reviewed that provides some insight on the range of interventions affecting learner performance is figure 3 below by Kremer, Brannen and Glennerster (2013) which provides comparative information on the cost-effectiveness of 30 primary school

RCTs aimed at improving learner performance measured through test scores. The left panel in the figure shows the estimated impacts measured in standard deviations (SD) at the 90% confidence interval. The right panel shows the SD gains in test scores per $100 spent.

In conclusion international literature shows that several initiatives based on improving access and participation are effective but improving learner performance is a greater challenge. Providing additional inputs has a limited effect when pedagogy and classroom practice are not strengthened simultaneously. However, the ideal combination affirmed in the literature is a combination of resources, accountability reforms and pedagogical support. The literature also points to differentiation in gains by gender as seen in PROGRESSA. It seems that streaming or adapting teaching to cater for differences in performance may therefore be more effective for improving learning. More research is required on how to best address this and the individual pacing and learning opportunities created by the availability of technology should form part of the options considered.

Figure 3: Test score gains compared to cost-effectiveness in 30 RCTs



Source: Brannen and Glennerster,2013

3.3 International Literature on Reading RCTs

The first reading study reviewed is an RCT on textbook provisioning in Kenya (Glewwe, Kremer and Moulin, 2007). Within the Kenyan schooling context mastery of English is important for secondary school entrance; in Grade 8 at the end of primary school learners write a national examination, the Kenya Certificate of Primary Education (KCPE). Only learners that pass this examination may proceed to secondary school. In response to the pressure created by the high-stakes Grade 8 examination, schools unofficially introduce culling with the highest dropout and repetition seen in Grade 7.

The RCT was implemented in a sample of 100 Kenyan schools in similar geographic locations, and with similar learner enrolment and pre-intervention test scores. Textbooks were provided to 25 of these schools each year from 1996 to 2000 for Grade 3 to 8. The textbooks were provided in English which is the medium of instruction in Kenyan schools although it is often a third language for learners. Textbooks were provided for English in Grade 3 to 5; Mathematics in Grade 3, 5 and 7; and Science in Grade 8.

On average the provision of the textbooks did not raise the overall average performance of learners, and neither selection nor attrition bias seems to have influenced the results. Only 16% of the median learners in Grade 3 classrooms and 28% in Grade 4 classrooms could read the textbooks, out of 50 schools selected randomly within the sample. This may be why the study only found improved performance for learners that had started the year with higher learning levels - there was a highly significant positive correlation based on the interaction of pre-test scores and the intervention. The authors' also point to the probability of teachers and schools privileging stronger learners through the manner in which teaching is delivered, particularly as there are high-stakes emanating from the primary school exit examination. This is supported by the positive prediction to proceed to secondary school for learners that benefitted from the textbooks. The findings also point to the effect of home SES even amongst the subset of rural schools. A 1990 Ministry of Education survey showed a pupil-textbook ratio of 17 to 1; similarly at the start of the RCT approximately 80% of the sample had less than 1 textbook to share. However, the stronger learners often tended to be those who had their own textbooks for some subjects. These findings were further

supported by a semi-independent school grant provided by the state to 25 schools within the sample. In the schools where the grant was used for textbook provisioning, the effect was only positive for stronger learners.

Considering the provisioning patters aligned with SES prevalent in the Kenyan schools coupled with the high-stakes primary school exit examinations, the Matthew Effect observed does not imply that textbooks do not contribute to learning. Rather, as concluded in the study, the findings reflect the gearing of the education system towards academically stronger learners; and the disadvantages of whole class teaching rather than streaming by ability. This reality is not only valid in the Kenyan context but similarly in the South African context. A qualitative component to the study, particularly in the initial years, could have contributed significantly to a greater understanding of the nature of classroom engagements and provided insights for revision of the design in the latter implementation. Such insights may furthermore contribute to knowledge on bridging the attainment differences within classrooms.

Differential impact also emerged from RCT data on the introduction of a remedial programme in India, where learners were matched with teachers according to their ability level. Learners gained 3.01 standard deviations in their test scores per $100 spent. Another study in Kenya, where an extra teacher was hired and the Grade 1 class was divided into two, based on initial learner performance, showed significant gains in learner performance (J-Pal,2012). There is a clear Matthew effect emerging from these studies- the notion that initially better-performing children typically gain more from additional interventions and from schooling itself. These differences in learner ability point to a need for differentiated interventions to address different magnitudes of learning gaps within schools. This is easier said than done.

In light of the emerging Matthew Effect identified broadly in the international literature on effective learner attainment and specifically in the study above, the second study reviewed (He, Linden and MacLeod, 2009) evaluates the relative efficacy of a literacy skills

development program on three different schooling populations. The Pratham Shishuvachan[4] was implemented in Mumbai in three different schooling institutions. During the first year 2679 learners in 67 municipal schools were randomly assigned to three research groups. The first group, consisting of 24 schools, implemented the programme as an afterschool activity. The second group, consisting of 23 schools, implemented the programme as part of the normal school programme. The third group, consisting of 20 schools, were the control group. In the second year a third treatment group based in the community and not attached to any educational institution was added, the econometric methods used to determine and equate the group to the initial treatment groups is discussed in the main report. The purpose of the Shishuvachan was to improve reading comprehension for four to five year old children prior to starting primary school. The programme had three main component, firstly, teacher mediated classroom-based exercises completed through the use of stories, flashcards, and posters; detailed lessons plans complemented by constant monitoring, training and coaching; and lastly, a community child library with age-appropriate materials.

Implementation took place over three years and an analysis of the findings shows that the programme was effective in all the different institutions. Learners in treatment groups gained between 0.12 to 0.70 of a standard deviation on their basic literacy test score, compared to the control group. The largest gains were in the afterschool implementation group, learners gained 0.55 standard deviations at the end of the first year compared to the 0.26 standard deviation gained by the in-school group. The programme was therefore most effective when it complemented normal teaching rather than replaced it. Learners made the largest gains in the first year of implementation, with smaller gains when learners started schooling in the second and third year of implementation. The programme was also most effective when implemented in pre-school, in line with the original design, and the category of learners that made the strongest gains in all three years were those who began at the lowest baseline reading score (He, Linden and MacLeod, 2009).

[4] Shishuvachan is a program conducted by the Pratham Mumbai Education Initiative, an educational NGO founded in 1994 in the slum communities of Mumbai. It was originally designed to supplement balwadi (daycare) programs run by Pratham, the government school system, and other organizations (He, Linden and MacLeod, 2009).

The success of the programme in improving the literacy performance of all learners, particularly prior to formal schooling supports the notion that early intervention is the most effective. The largest gains which were made by learners that had the lowest performance in the baseline may point to the appropriateness of the initial design as the programme was intended to provide basic competencies in preparation for schooling based on the low levels of literacy amongst the population and children prior to schooling. The literature, including the Kenyan study reviewed above, seems to indicate that curriculum expectations increase exponentially as schooling progresses and this initial grade-appropriateness is not maintained beyond pre-primary schooling. A further component on the largest gains was programme effectiveness as an after-school intervention. This may be reflective of the structural advantages that the schooling system provides such as infrastructure, and accountability through standard school management. The ability of the programmes to leverage these functional components of the schooling structure complemented with providing training and resources to volunteers who implemented the programme points to the strength of providing accountability, structure and resources.

One of the possible reasons for factors identified in the systematic literature review was that resources were most effective when coupled with improved accountability and capacity. This seems to be supported in this study as the largest gains were made when the programme was implementation through pre-schools, an existing education institution with standard accountability practices in addition to the programme specific resources and training.

The third international reading RCT examined is the Banerjee et al. (2010) community participation study, implemented in Uttar Pradesh (UP), India's most populous state. The government had established Village Education Committees (VECs) in the early 2000's to encourage community engagement. The VECs consisted of three parents, the head-teacher of the village school, and the head of the village government. Their responsibilities included hiring additional teachers, allocating school resources and holding teachers accountable by monitoring learner performance. By 2005 the performance of learners in UP had worsened and there was a sense that these structures had become ineffective. This study was premised on the hypothesis that community participation can improve learner reading

outcomes even in poor communities. The three components comprising participation were: easy to understand information on the current levels of learner performance, the provision of monitoring tools and training to enable communities to monitor learner performance consistently and independently, and the activation of VECs for collective action. Three interventions following the logic of the TOC were developed and implemented in 280 villages in the Jaunpur district of UP. Each intervention group consisted of 65 villages. The remaining 85 schools became the control group. The first intervention comprised of information only, facilitated by Pratham[5], with a focus on highlighting learner performance and the role of the VECs. The second intervention entailed the provision of information, as well as training community volunteers to administer a short reading test and compile a simple report card. These reports were presented in village-wide meetings. In the third intervention, intervention two was delivered by Pratham and community volunteers. This was complemented by pedagogical training on basic techniques to teach reading. The volunteers held voluntary reading camps after school that lasted two to three month. More than 400 reading camps were established in over 55 villages.

All of the interventions were implemented fairly successfully with an average of 100 community members out of 360 in a village, attending the meeting held. The baseline was conducted from March to April 2005 and the endline a year later. One of the main findings is that the interventions did not increase community participation in governance and management of schools. Secondly, none of the interventions had a significant improvement on learner performance except in treatment three. Learners in treatment three were 1.7 percentage points more likely to read at least letters and 1.8 percentage points more likely to read words or paragraphs (significant at the 95% confidence interval). When using learner attendance at the reading camps as an IV, this estimate becomes substantially larger. On average, attendance resulted in learners being 22 percentage points more likely to be able to read at least letters, 23 percentage points more likely to read at least a word, and 22 percentage points more likely to read a story (the first two estimates are significant at the 95% confidence level). An analysis based on heterogeneity showed that learners who could

[5] Pratham is India's largest education NGO. They led this study involving researchers from the World Bank and Abdul Latif Jameel Poverty Action Lab (J-PAL) at the Massachusetts Institute of Technology (MIT) Banerjee et al. (2010).

not read at all made the largest gains, they were 60 percentage points more likely to be able to recognise letters after a year than their counterparts in the control villages. Learners that could only read letters at the baseline were 26 percentage points more likely to read stories, and learners that could read words or paragraphs were 46 percentage points more likely to read stories (the first two estimates are significant at the 95% confidence level).

One of the main findings from the literature earlier in this chapter, was that learner participation alone does not result in effective learning. This study highlights that this also applies to community and parent participation. The establishment of school governance bodies, endorsed by the state does not guarantee effectiveness and it seems that coordinating collective action is not the binding constraint in community participation. Information and opportunities for participation alone, seem to be ineffective in affecting learning and sustained community engagement. Community members seemed more committed in providing direct support to learners that showed commitment to learning. The establishment of reading clubs by volunteers was maintained and successfully in changing learning outcomes with educationally meaningful and statistically significant gains for those that attended. This supports the assertion that improved accountability, through the information shared, coupled with pedagogical knowledge and resources can change learner outcomes. In this case this was transacted with the volunteers rather than through teachers in the school, there may be interesting findings if the same model is implemented with teachers. The difference in learner ability have again, been highlighted in the study. The authors indicate that the programme was geared towards non-readers and was therefore successful as this sub-category gained the most. It is still interesting to note that depending on initial competencies, learners mastered the next level of complexity but none of the learners improved in every sub-category.

The fourth and final reading RCT reviewed is a one-month “read-a-thon” program implemented in the Philippines by Abeberese, Kumler and Linden (2011). The RCT was implemented in Tarlac province in the Philippines, in a sample of 100 schools, with 50 as the intervention group and 50 as the comparison group. The read-a-thon took place between September and November 2009, with learners assessed immediately after and again in February 2010. Each read-a-thon school received 60 grade appropriate books in English and

Filipino, targeted at Grade 4 learners. A reading specialist from the implement agent, Sa Aklat Sisikat (SAS) provided teachers with training on conducting a read-a-thon and this entailed encouraging children to read as many as 60 books within the next month. A tracker on the books read was also developed and distributed.

The initial assessment results following the read-a-thon showed gains for the treatment group who scored 0.13 standard deviations more than learners in the comparison group in reading proficiency tests. However, the gains made declined over time, with treatment group learners scoring 0.06 standard deviations more than the comparison group in the follow-up assessment several months later. Learners in the treatment group continued to report reading more books during the read-a-thon and after. The policy lesson from this study is that although short interventions may be useful in creating an interest with some gains made in learner performance, maintaining these gains requires sustained support, monitoring and accountability. This again supports the primary finding from the literature review that improvements in learner performance require resources, changes in pedagogical practice and accountability.

3.4 National Literature on RCTs in Education

There are a very limited number of RCTs that have been completed in South Africa on education interventions broadly or reading. Using the 3ie and JPAL evidence mapping tools, only two studies appeared and one was quasi-experimental. Resultantly, only three studies will be discussed in this section with only one study focusing on reading. These are the Mind the gap study guide RCT (Department of Basic Education, 2013), a computer-based maths RCT by Böhmer (2014) and the Systematic Method for Reading Success (SMRS) by Piper (2009).

There are two main studies referenced in the work by Snilstveit et al. (2015) on participation and access, which are based on the analysis of existing national databases rather than interventions designed and evaluated nationally. These studies show an impact on the enrolment figures nationally due to the introduction of the no-fee policy in government

schools based on SES and child grants. These studies will not be discussed further as there is near universal access to basic education in South Africa.

With regard to improving learner performance, there are a handful of RCTs published in South Africa. The first of the two recent RCTs that will be discussed in this section is the Mind the Gap study guides RCT. The study focused on the development and distribution of study guides in an effort to assist Grade 12 learners to 'mind-the-gap' between failing and passing, by 'bridging-the-gap' in learners' understanding of commonly tested in the National Senior Certificate (NSC) examinations. The RCT, conducted in 2012, measured the impact of the study guides on performance in the NSC. Initial distribution of the study guides was not at scale nationally but limited to underperforming districts in the Eastern Cape, Limpopo and Northern Cape. This provided the opportunity for an RCT which was implemented in Mpumalanga. The final sample for the RCT consisted of 318 ordinary public schools, with 79 schools receiving the study guides and 239 schools allocated to the control group. Distribution took place in September 2012, approximately 6 weeks prior to the NSC exams, and analysis of the NSC results indicates statistically significant positive impacts on learner performance for schools that received the study guides in Life Sciences and Geography.

The results indicate that learners who received the study guides in Geography scored 1.9 percentage points more than they would have had they not received the guides. For Life Sciences the impact was 2.17 percentage points but results in Accounting and Economics did not indicate statistically significant differences between learners who received the study guide and those that did not. A simulation exercise using NSC 2010 data demonstrates that 5609 children who did not pass matric in 2010 would have passed matric had the Geography and Life Sciences study guides been distributed nationally. Following the RCT the study guides were distributed nationally and are available online (Taylor and Watson, 2013).

Discussions on the possible causes for a lack of improvement from the Accounting and Economics study guides suggest that these subjects require extended practise and the timing of the study guides was not aligned to this requirement; also the Geography and Life Sciences study guides may have been of better quality and provided material otherwise not easily accessible to learners; such as mapwork materials in the case of Geography.

The second study reviewed is by Böhmer (2014). The study examines the role of computers and technology in assisting schools in delivering learning to a class with a broad range of abilities, one of the key areas emerging from the international literature review as an area for further research. The RCT was based on a programme by Numeric, an NGO that provides an after-school mathematics intervention focused on improving numeracy levels using computer assisted learning. The intervention took place after-school during the week or on a Saturday morning with learners attending Numeric classes for an hour and a half on a bi-weekly basis. The classes consisted of approximately 20 learners and were facilitated by coaches who were mostly trainee maths teachers. The content of the classes mostly consisted of arithmetic drills that got more complex over time. Immediate feedback was provided to learners and the pace and content was based on learner ability rather than curriculum specific exercises or pacing.

The sample consisted of 9 schools in the Cape Town metropolitan area with a total of 2008 learners. In each school, 40 learners per class were selected for participation, with a total of 472 learners in the intervention group. The schools that finally participated had to apply through a two stage process and the criteria for exclusion included poor access to computers, likelihood to drop out and learners that had to travel long distances to school. The final sample of participating schools consisted of urban, well managed schools with working computer labs and an internet connection. The programme was implemented over one year with the endline conducted in November 2014. The basic numeracy scores of participants were 0.321 standard deviations higher than the comparison group. There was a similar difference in the provincial assessments, with the treatment group scoring an average of 0.246 standard deviations above those in the control group.

In interpreting the findings, it is difficult to measure the true impact of the programme as coaching in the programme was delivered by university maths students who could arguably produce the same results using ordinary teaching methods such as pen and paper. Secondly, there seems to be selection bias in determining the sample and there are issues of external validity considering the niche characteristics of the final sample. The researcher acknowledges that the treatments schools were not representative of the larger population

nationally or provincially and that the results may also reflect school factors rather than just the intervention. However, this study is a good example of within-school randomization, hence the ability to have only a few schools in the sample. This should be a considered option when implementing interventions locally.

These findings from general curriculum-based RCTs provide insights for interventions focused on reading and literacy especially as there are a small number of RCTs focused on reading interventions locally. Some studies ostensibly designed with a control group, such as that of Pretorius (2014), are case studies and do not truly fulfil the criteria of RCTs since they fail to have sufficient sample size or did not actually randomise treatment assignment. One study which comes close to a fully-fledged RCT is the evaluation of the impact of the Systematic Method for Reading Success (SMRS) by Piper (2009).[6] This was conducted in Mpumalanga, North West and Limpopo using the Early Grade Reading Assessment for assessment. There were 546 learners at the Grade 1 level in the sample with 10 treatment and 5 comparison schools per province and approximately 10 or 20 learners assessed per school depending on the number of classes.

A total of 21 out of a possible 45 SMRS lessons were implemented on average between the baseline data collection in February 2009 and the endline in June 2009. The assessment focused on four tasks: letter sound recognition, word recognition, reading a simple passage and comprehension questions. The impact of the programme was positive, showing improvements in the various assessment areas. Learners showed gains of 4.56 words per minute on average which may be measured as 0.79 standard deviations. The average words read correct in a passage also increased by 7.21 words or 0.80 standard deviations. Similarly in reading, learners in the treatment group scored 8.24 percentage points higher than the comparison group. The programme was premised on closely following the SMRS lessons which systematically teaches language focusing on the various components within languages. Despite a few study limitations, this evaluation pointed to the possible success of a programme using a systematic instructional method for teaching reading. The Early Grade

[6] One limitation of the Piper (2009) study is that the sample was rather small for the South African education setting – 29 treatment and 15 control schools. There is also no mention in the report about whether standard errors were adjusted for the fact that learners were clustered in schools, something which is critically important but often neglected, at least in South African studies on school surveys.

Reading Study, to be described in detail later, can be regarded as a follow up to the preliminary evidence provided by the study on the SMRS programme.

3.5 Conclusion of Literature Review

In conclusion, the review of the international RCTs literature points to specific kinds of interventions that affect participation and access, but most importantly, improve learner performance. Materials that are structured, embedded in tested methodologies and theory and mediated through effective training and finally complimented with improved monitoring emerge as a promising combination, especially when the desired outcome is improved reading acquisition. The use of temporary teachers or coaches and the incentive systems underlying their performance are clearly areas for further research, with the understanding that there are different dynamics at play and thus external validity is a valid concern. The literature also clearly points to the usefulness and positive impact of grouping learners by ability and providing them with differentiated support. The practicality of this without the availability of additional resources or support through programmes that may include technology is still an area requiring research especially in the South African context. It is also clear, from the South African literature, that there is a knowledge gap across the board in the use of quantitative methods for curriculum research and specifically RCTs. Where there has been an attempt to use these methods the understanding of the technical requirements of implementation and measurement sometimes result in compromised work that is difficult to truly interpret as quasi-experimental or experimental. As researchers in education and policy makers engage and develop an understanding of the methodologies, the use and utility of these would become more available.

CHAPTER 4: CASE STUDY 1: READING CATCH –UP PROGRAMME

4.1 Introduction

Critical requirements in research intended to inform policy decisions are firstly, the relevance of interventions and findings for the larger schooling population and secondly, the precision of the measured effect. The methodological critique that is levelled against much of the research and evaluations in education nationally has been critically discussed in Chapter Two. The primary contribution of internally valid methods, specifically RCTs, is the ability to provide rigorous and valid estimates responding to both of the critical requirements with high levels of certainty. This is achieved through large sample sizes that may be designed to be nationally and provincially representative of the schooling population, as well as random selection of schools or learners in evaluating interventions. Based on these parameters, it is possible to infer findings from a sample to the larger population with the expectation that there are no reasons to expect systematic differences between participants and non-participants.

The gaps in existing knowledge and limitations in the methods used by existing South African studies have been clearly demonstrated in Chapter Three. This chapter provides empirical evidence supporting the theoretical and methodological significance of using RCTs in education research, specifically for the South African context, by illustrating the policy relevance and contribution that is possible through RCTs. It demonstrates the value of RCTs through a secondary data analysis of a South African Reading Catch-Up Programme (RCUP) RCT undertaken in Pinetown, KwaZulu-Natal (KZN). The chapter begins with a discussion of the RCUP study's research design, including sampling, the data-collection processes and other methodological aspects. This is followed by a discussion of the main findings from the RCUP which have been reproduced for this analysis. Chapter Five provides an analysis of data from the second RCT case study, the Department of Basic Education's Early Grade Reading Study (EGRS).

The RCT methodology provides an opportunity to measure heterogeneous effects empirically- establishing if the impact of a programme differs depending on various learner,

school or teacher characteristics. The second section in the chapter explores this through a description of the interactions between learner characteristics, including gender, and the extent of the Matthew Effect - the notion that initially better-performing children typically gain more from additional interventions and from schooling itself. This section concludes with an analysis of learner characteristics most associated with the Matthew Effect. In light of the Matthew Effect and associated learner characteristics, Chapter Five provides a descriptive analysis of the EGRS data and identifies the prevalence of similar learner characteristics at the very start of formal schooling.

A multivariate regression method is used for the analysis through the statistical programme STATA. Random assignment to the treatment group (in combination with a large enough sample) allows one to assume that there is a balance in all observable and unobservable factors across the two groups of schools. There is no reason to expect that one set of characteristics such as parental education, or SES, is different in either the intervention or control groups in both case studies (Taylor, 2014). Nevertheless, in order to increase the explanatory power of the regression model, multivariate analysis is conducted to control for certain observable characteristics of the sample, such as SES, gender and age. This analysis, as a result of internal validity within the RCT methodology, provides statistical estimates of correlations between learner performance and learner characteristics. Most notably, the correlation between treatment assignment and outcomes of interest can be interpreted as the causal effect of the intervention.

4.2 Background

The RCUP was developed to strengthen the English skills of children in grades 4 and 5, whose first language is not English but who are required by national education policy to learn using English as the language of instruction from Grade 4 onwards (having used their home language during their preceding years of schooling). The development and initial implementation of the RCUP was conducted as the Gauteng Primary Language and Mathematics Strategy (GPLMS) between 2011 and 2014. A preliminary evaluation of the GPLMS indicated large gains in the language skills of programme recipients over the intervention period (Hellman, 2012). However, there was no comparison group. Therefore,

some strong assumptions had to be made about how much learning would have taken place over the period had there been no intervention. The inadequacy of using Hellman's approach to estimate impact has been elaborated on earlier in this study, suffice to say that the method used cannot estimate impact in a scientifically rigorous manner addressing both the criteria of internal validity and selection bias.

In view of the promising yet inconclusive findings of the initial evaluation, a team of independent researchers (Fleisch, Taylor, Schöer and Mabogoane: 2015) undertook a more rigorous impact evaluation of the RCUP in the district of Pinetown in the KwaZulu-Natal province in 2014. Considering the initial estimates of high impact based on the evaluation by Hellman using non-experimental methods, the decision to conduct an RCT of the same intervention is significant. The RCT would provide causal estimates of the impact of the intervention, resulting in high levels of confidence in making policy recommendations about the effect size. In light of the possibilities of scalability to other provinces, the contribution of the RCT in informing the investment decisions of donors, policy makers and other education stakeholders is immeasurable.

4.3 Intervention Design, Sample, and Data Collection

The hypothesis underlying the Pinetown RCUP programme was that the learning gaps in English at the end of the second term in Grade 4 might be at least partially made up through the provision of a well-designed relatively short intervention. The RCT comprised 40 intervention schools with a comparison group of 60 schools. The intervention lasted for 11 weeks; April to June 2014; and consisted of on-site teacher support by reading coaches, the provision of scripted lesson plans and additional graded reading books. The sample consisted of 100 underperforming ordinary public English LOLT schools, derived from Quintiles 2 to 4 schools which had scored 55% or below on the Grade 4 First Additional Language (FAL) test in both the 2012 and 2013 Annual National Assessment (ANA) tests. The schools selected also had to have entered between 15 and 120 learners in the 2013 ANA test for Grade 4 FAL (Fleisch, Taylor, Schöer & Mabogoane, 2015).

Some of the additional assumptions were an 80% power level, and a 5% significance level with the Intra-Class Correlation Coefficient value (between school variance as a proportion of total variance) assumed to be 0.20. Under these assumptions, the study was powered to identify a minimum detectable effect size of 0.15 standard deviations, which equated to 3.5 percentage points in the test for the study. As the study took place over 11 weeks during the middle of the school year, with the pre-and-post-tests extending this to 12 weeks, attrition did not pose a concern (Fleisch, Taylor, Schöer & Mabogoane, 2015).

All Grades 4 learners in the schools participated, as the intervention was targeted at the classroom level, and administered across the classrooms rather than to individual learners. The unit of analysis was therefore the school, but learner data was collected. Although 100 schools were targets, the final sample consisted of 96 schools: three schools in the control group were replaced by schools selected by the district office and having not been randomly selected, were excluded in the analysis. One other school was effectively excluded since it participated in the post-test but not in the pre-test – since the regression models include pre-test scores, this effectively renders the entire school missing. The data obtained on the pre-test was for 2663 learners from 96 schools but for analysis purposes, only data from the 2543 learners who also wrote the post-test was used.

The programme had several different learning resources for the classroom. These were printed lesson plans covering 70 lessons; two A4 learner exercise books for each learner, one to write in during regular class time, and a second specifically for tests; four listening and speaking posters that covered four themes; a set of 12 graded reading books; a set of 12 reading sheets with 'look and say' words; and an assessment record book. The intervention implementation was mediated by in-class coaches who trained teachers on the materials and conducted classroom-based support including modelling lessons.

The Theory of Change (TOC) is based on literature on system-wide change noting however that the most significant factor influencing learning is instructional practice mediated through teachers. The delivery of instructional practice is however dependent on the effectiveness of teaching enables by the use of learning resources in a structured manner with access to support, both in pedagogy and content knowledge. In the RCUP, this shift in

instructional practice was seen as integral component enacted through the delivery of a coherent set of materials including new lesson plans aligned with the curriculum as well as other curriculum support resources mentioned. The most notable component of the lesson plans is the provision of new instructional practice which includes faster paced instruction, improved curriculum sequencing, and the embedding of core methodologies within the materials with opportunities to practice these. The role of the reading coaches was both as a capacity building measure and as a model of accountability anchored in learner attainment and support, the significance of this relationship has been explained in the preceding chapters.

The assessment component consisted of teacher questionnaires administered at the beginning and at the end of the study, and learner pre- and post-tests. The learner baseline test contained 35 items in four categories, namely, spelling (20 items), comprehension (6 items), language or grammar (6 items) and writing (3 items). The endline test was exactly the same as the baseline text except for the addition of two spelling items and two comprehension items. These were relatively easy items added in response to the floor effects noted in the analysis of the baseline test, in order to improve the distribution of scores in the endline and thus improve the chances of picking up the treatment effect. Different organisations were contracted to implement the intervention and to conduct the data collection for evaluation. The evaluation agent was kept blind to which schools were in the intervention group versus the comparison group.

What is significant about the research design of the RCUP, in-line with RCT methodology is the large sample which enables the estimation of the impact of the interventions with sufficient statistical precision; an upfront determination when selecting a sample, including the size of the intervention group and comparison group. A letter of permission from the Principal Investigator for the use of the RCUP is attached as Appendix B. Detailed information on the sampling process, the data collection, the intervention and assessment components are available in the main report [7]by the researchers.

[7] The researchers attained permission to conduct the study through the KwaZulu-Natal Department of Basic Education. The main report for the study is attached as Appendix C.

The researchers attained permission to conduct the study through the KwaZulu-Natal Department of Basic Education. Furthermore, a pre-analysis plan [8]was published on the RCT registry of the American Economic Association (https://www.socialscienceregistry.org/trials/405) in order to address any concerns of data mining, an unhealthy practice in the field which was discussed in Chapter Two, as well as reduce the chances of publication bias if there were negative results.

4.4 Learner Characteristics

This section provides descriptive information on learner characteristics based on data collected in the study. Although not strictly designed to be representative of any wider population, in a rough sense this sample of schools is not altogether unusual in any way and is therefore of broader interest (DBE, 2015).

Table 1: Learner characteristics by gender and age

Learner Age	No. of Female Learners	Cumulative % of Female Learners	No. of Male Learners	Cumulative % of Male Learner	Total No. of Learners	Cumulative % of all Learner
8	7	0.56%	4	0.31%	11	0.43%
9	636	51.6%	436	34.27%	1,072	42.81%
10	376	81.78%	456	69.78%	832	75.69%
11	99	89.73%	225	87.31%	324	88.5%
12	23	91.57%	74	93.07%	97	92.33%
13	11	92.46%	35	95.79%	46	94.15%
14	94	100%	54	100%	148	100%
Total	**1,246**		**1,284**		**2,530**	

Source: Replication of analysis from RCUP data. The age groups indicated in the table above exclude data for learners that reported to be younger than 8 or older than 14.

The modal age for learners in Grade 4 is 9 years old, which is in line with the expected learner age based on the prescription of compulsory schooling being from the age of 7 to 15 years old, with learners starting Grade 1 at the age of 7. However, approximately 82% of girls were aged 10 or younger while only 70% of boys fell into the same age category. This reflects interesting gender patterns already evident at the beginning of the Intermediate Phase. The reasons for this are not evident from this data; however, it is reasonable to

[8] The pre-analysis plan is attached as Appendix D.

expect these patterns to point to education system inefficiency such as repetition, with males having a higher probability of repeating (Department of Basic Education 2013). The overall proportion of over-age learners, calculated as learners aged 12 and above, was 11.5%, fairly high but not unusual for South Africa. Based on the average age in this sample and by national standards, at the age of 12 learners should be in Grade 7 or in the worst case, in Grade 6. This figure of 10.27% for female learners and 12.69% for males similarly reflects differences in learner progression by gender.

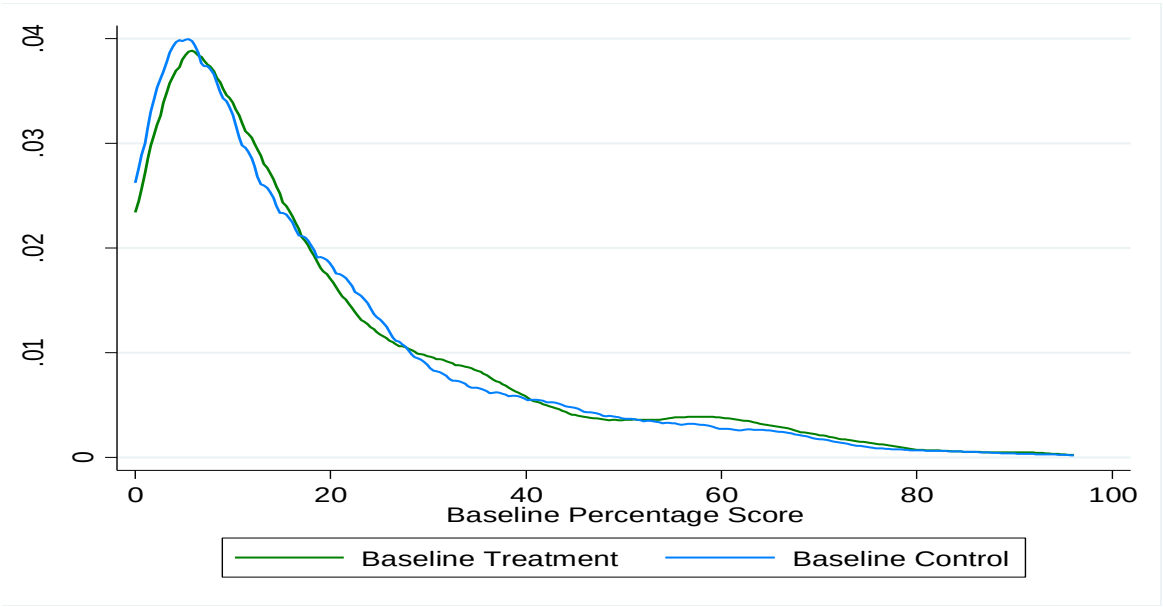
Table 2: Learner characteristics by location and school quintile

School Locality	Quintile 2	Quintile 3	Quintile 4	Total Number of Learners	Percentage of Learners
Rural	78	792	351	1,221	49.5%
To be updated	0	98	0	98	4%
Urban	0	488	659	1,147	46.5%
Total	**78**	**1,378**	**1,010**	**2,466**	

Source: Replication of analysis from RCUP data

The Pinetown district is a 1504 square kilometre (km) district described as mostly urban but including informal, and traditional rural settlements. The average distance between schools and the PED District Office is 17.4 kms (DBE, 2015). The schools in the sample are almost equally categorised into rural and urban, with Quintile 3 being the largest quintile at 55.88% and containing 792 rural schools. According to DBE data there are only 5 Quintile 1 schools, constituting 1% of schools in the district; as such, the exclusion of Quintile 1 schools from the research sample is warranted. The broad representation of the different quintiles in the district and the overall large sample means that there is a higher representivity of the larger population captured within the sample, a strength of the RCT methodology.

Figure 4: Baseline Performance by Intervention Group



Source: Replication of analysis from RCUP data.

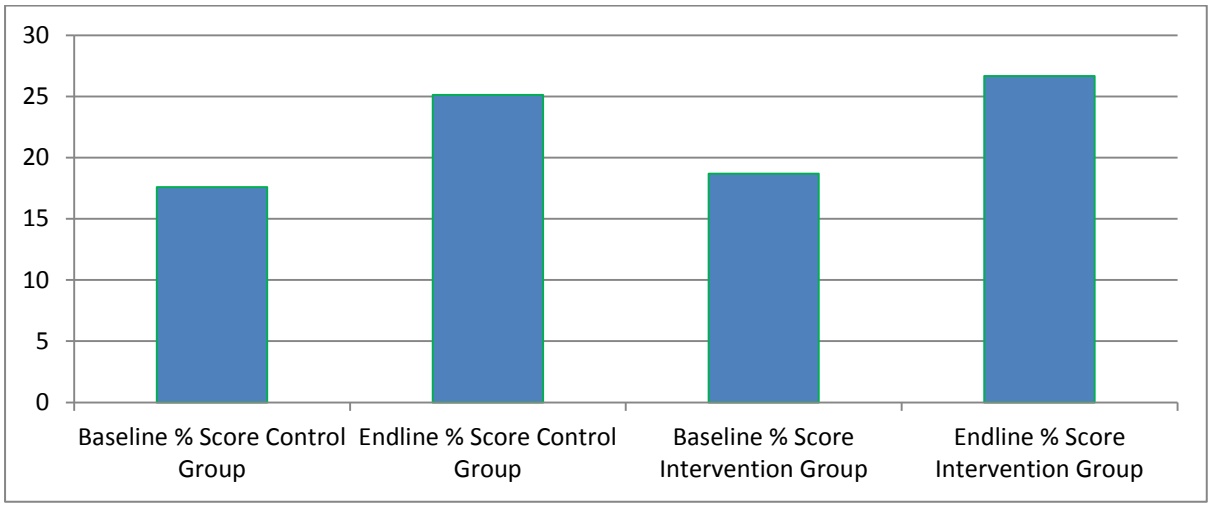
Learner performance was extremely poor as measured by the baseline test, with the majority scoring less than 20%. Although this test was designed by the researchers, it was based on the expected curriculum coverage as indicated in national curriculum standards and policies. The performance of learners clearly indicates large deficits in learner content knowledge by the start of the Intermediate Phase, as the test administered contained curriculum content from the Grade 1 to Grade 4 level.

The baseline test also points to test composition considerations that are a critical aspect of research based on learner assessment. There is an inherent tension between creating tests that are able to discriminate at the bottom-end of learner performance in order to categorise learner ability more accurately; and testing at the appropriate Grade levels based on curriculum expectations for the Grade. This tension is clearly illustrated in the baseline test as the distribution of learner performance points to 'floor effects'-when a testing instrument has a lower limit to the data that it can reliably specify (Groth-Marnat, 2009). There is merit in understanding the actual curriculum level of learner ability regardless of grade, but for the purposes of this study there was an imperative to test against the Grade 4 curriculum in order to measure the effect of the intervention which is pegged at the Grade 4

level. This consideration remains significant for general purposes of research and interpreting the results and findings from interventions.

The performance of learners in the baseline test, although weak, was similar for both the intervention and comparison group, affirming that the randomisation, a critical criterion of RCTs, was successful. There were no systematic differences between the groups and thus the changes noted in the performance of the intervention group may be attributed to the intervention.

Figure 5: Mean score for Intervention and Comparison Group for the baseline and endline test



Source: Replication of analysis from RCUP data.

Figure 5 clearly demonstrates that the intervention did not make a difference to the intervention group compared to normal schooling in the treatment group. On average the gains made by learners between the baseline and endline was 7.74 percentage points, although some learners scored zero in both the baseline and endline. Importantly, there was no significant difference between the gains in treatment group and those in the control group. The figure clearly illustrates the value of a counterfactual, resultant from using RCTs. There is no reason to expect the intervention and comparison group to be systematically different as a sufficiently large sample was used, and allocation to the groups was random. Without a valid comparison group a false positive impact may have been concluded.

The results of the RCUP evaluation should contribute to a sobering realisation that the effects of the various interventions introduced by education stakeholders including NGOs and government are not obviously positive or more importantly, different from normal schooling hence the need to evaluate programmes before they are rolled out provincially or nationally. This is a significant policy and research question that warrants the use of RCTs and other rigorous methods to evaluate programmes in education.

Encouragingly, there seems to be a noteworthy quantity of learning that happens in the third term in Grade 4 as seen in the gains made across both the intervention and comparison groups; however, these remain at a low base with the average endline score being approximately 25%. This raises questions about the quality of learning that happens in the Foundation Phase in English FAL. Considering that the Theory of Change for this intervention is based on changing the instructional core of teaching through the consistent and systematic use of lesson plans, it is possible that this 11-week intervention was simply too short to significantly influence classroom practice and learning. As such, the duration of interventions is an important factor to consider while noting the tension of the need to establish effective short remedial programmes. More research is definitely required in this area. There is room for careful consideration of the reasons for no impact, the reasons mentioned in this discussion are not exhaustive.

Table 3: Learner performance by gender

Test Type	Number of Male Learners	Average test score for males%	Standard Deviation	Number of female Learners	Average test score for females%	Standard Deviation
Baseline Test	1293	14.89	16.30	1250	22.03	19.71
Endline Test	2543	21.67	19.70	1250	30.76	22.89

Source: Replication of analysis from RCUP data

Females performed better than males in both the baseline and endline tests but both males and females made similar gains between the two tests, as illustrated in Table 3. At the baseline there was a difference of 7.14 percentage points between males and females, with the gap increasing to approximately 9.09 percentage points in the endline test. According to academic literature on gender in South Africa and according to virtually all local and international assessments that South Africa participates in, girls are outperforming boys,

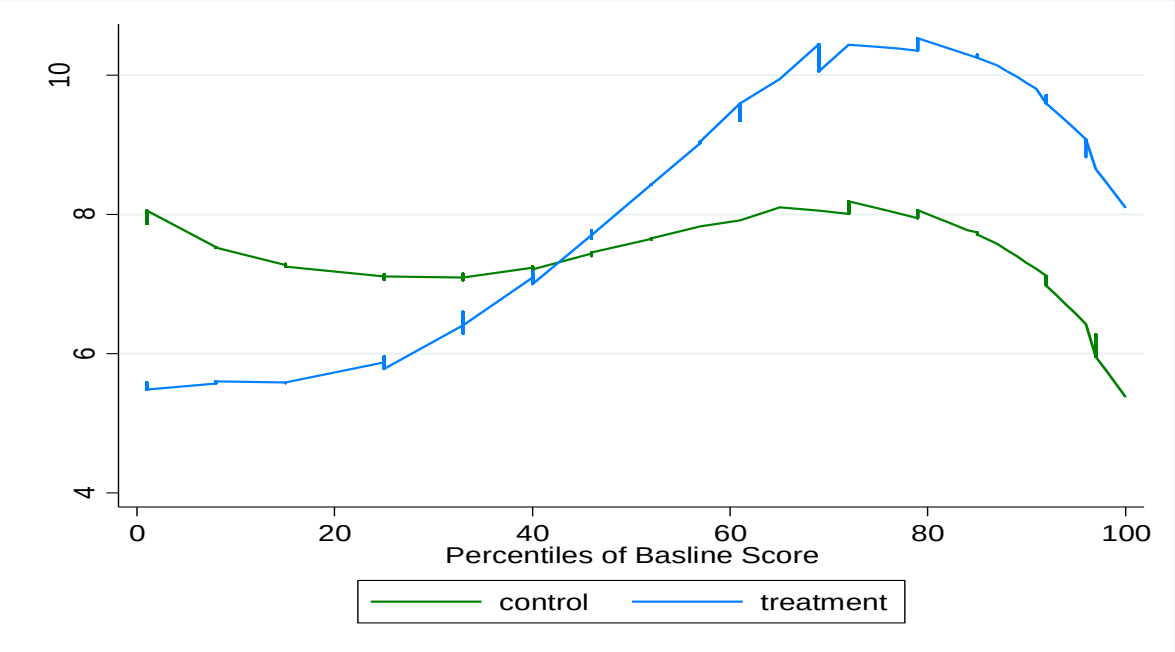
especially in reading and literacy. Perhaps as a consequence, boys have higher grade repetition and drop-out rates (Zuze and Reddy, 2014). This data thus provides insight into patterns of learner performance over and above the differences between treatment and control groups. The next section, for example, illustrates how the RCT data can shed light on the differential impact of interventions depending on the initial proficiency of children.

4.5 Learner Performance

In the main RCUP report heterogeneous effects were examined, namely, if the impact of the programme was different depending on various learner, school or teacher characteristics. Preliminary data analysis points to a correlation between the gain score attained and initial performance. There was no evidence of heterogeneous effects based on learner gender, age or exposure to English. This section will further explore the hypothesis that there are differences in learner performance based on their prior knowledge, whether there is a Matthew Effect.

Figure 6 provides a summary of the average 'gain score' based on 100 baseline performance percentiles. This allows us to examine the relationship between the baseline percentile and the learner gain score in percentage points using a Lowess smoothing line. A Lowess smoothing line is produced through a Lowess regression model, which is a non-parametric regression method that carries out multiple locally weighted regressions and then smooths over the distribution. Figure 6 shows Lowess smoothing lines separately for intervention and control groups.

Figure 6: Average gain score for the Intervention Group and Comparison Group



Source: Own calculations from RCUP.

The figure clearly illustrates a 'Matthew Effect' the graph shows a clear break at the 40% mark with increasing gains made by the Intervention Group, peaking at the 80% mark. The performance pattern prior to the 40% mark seems to be inversely correlated, with the Comparison Group making the largest gains. The caveat in interpreting this graph is that the majority of learners performed below the 40% mark and this should be kept in mind when interpreting the graph.

This figure demonstrates that RCTs go a lot further than simply providing a kosher stamp for evaluations, as in the USA, a discussion provided in Chapter Two. This analysis rather illustrates the insights, possibilities and patterns of performance which are possible to establish using the richness of RCT data. This new insight from the RCUP data, is only possible because of the sample size and the comparison group which are key RCT methodological features. One of the possible explanations for these differences supports the hypothesis that learners with a better grasp of the basics in the curriculum make the most gains in terms of catching-up curriculum-based interventions. Despite the intention that the RCUP would cover topics from the curricula of earlier grades, it may have actually been pitched at a higher level than the lessons implemented in the control group. At the

same time, the change in the pace and level of the curriculum for learners that are far behind may have resulted in even larger gaps as they are not able to make the leap between what they know and the new content.

The hypothesis about the Matthew Effect and extrapolations about possible classroom practice are differently illustrated in the table below, which shows average gains for each decile of baseline scores. The data in the table is organised into larger clusters which better illustrate the two inverse trends highlighted in Figure 6. The table clearly illustrates the two subsets of learners in classrooms in the sample as well as how different approaches to learning may impact on these groups, as illustrated in the inverse relationships seen between the gains in the Intervention Group and Comparison Group respectively.

Table 4: Decile average gain score for the Intervention Group and Comparison Group

	Mean of 'Gain Score'		
10 Deciles of Baseline Score	**Treatment Group**	**Comparison Group**	**Total**
1	7.39	5.65	6.70
2	7.48	5.71	6.74
3	7.06	5.55	6.50
4	6.77	5.31	6.11
5	7.15	7.56	7.33
6	7.61	9.78	8.60
7	8.01	9.83	8.71
8	9.59	13.00	10.96
9	8.97	10.79	9.79
10	5.42	8.61	6.89

Source: Own calculations from RCUP data.

The tables that follow display analysis from regressions on the extent of the Matthew Effect based on various test sub-components based on a categorisation by performance.

The main regression (1) in Table 5 shows the overall outcome variable, namely the post-test; while the remaining three regressions show the performance of learners based on clusters of performance in the baseline score, namely, the poorest performing 33%, those performing between 34% to 66%, and finally those performing above 67%. All of the regression models include controls for other observable characteristics as specified below, while the baseline accounts for prior knowledge.

Table 5: Main regressions

	(1) Full sample	(2) Bottom 33%	(3) Middle 33%	(4) Top 34%
Treatment	0.71	-0.76	0.75	2.49*
Standard error	(0.68)	(0.71)	(0.97)	(1.43)
Observations	2466	982	717	767
R-squared	0.76	0.18	0.21	0.62

*Note *p<0.1 **p>0.05 ***p<0.01*

All models include controls for baseline score, stratification dummies and learner age dummies. Standard errors are adjusted for the fact that learners are clustered in schools.

The overall estimated effect of the intervention is 0.71 percentage points relative to the Comparison Group, although this estimate is not statistically significantly different from zero. There is a negative effect of treatment estimated for those in the bottom 33% in terms of baseline performance, though this estimate is also not statistically significant. For the middle 33% and the top 34% there is a positive treatment effect which is noticeably larger for the highest performing learners. For the top 34% the estimate is statistically different from zero at the 90% level of confidence. Although the sample size is reduced, this is arguably still noteworthy. The following table explores whether there is evidence of a Matthew Effect within each of the skill domains within the test.

The regressions in Table 6 reflect the treatment effect based on interaction with the baseline scores of learners. Overall, the intervention effect was largest for the subcomponent of Language with a treatment effect of 2.92 percentage points, an estimate that is statistically significant from zero at the 99% confidence interval. We thus have a high level of certainty that learners made the highest gains in this subcomponent. The most significant results for this study are the positive coefficient for the interaction between treatment and baseline for 3 out of the 4 test subcomponents. With the exception of the Writing subcomponent, there are additional gains based on the treatment interacting with the baseline score. The largest gains are in Spelling, where learners gained an additional 0.06 percentage points, an estimate that is significant at the 90% confidence interval. The interaction effect in Language, at 0.05 is also high although it is not statistically significant.

Table 6: Regressions of overall test and subcategories with intervention and baseline interaction

	(1) Overall	(2) Spelling	(3) Comprehension	(4) Writing	(5) Language
Treatment	-0.33	0.46	-1.89	0.87	2.92***
Standard error	(0.82)	(0.65)	(1.56)	(1.44)	(1.1)
Baseline percentage score	0.98***	0.93***	.0.79***	0.49***	0.67***
	(0.02)	(0.02)	(0.03)	(0.04)	(0.03)
Treatment x baseline	0.05	0.06*	0.03	-0.03	0.05
	(0.04)	(0.03)	(0.04)	(0.06)	(0.04)
Observations	2466	2466	2466	2466	2466
R-squared	0.77	0.77	0.52	0.27	0.45

*Note *p<0.1 **p>0.05 ***p<0.01*

All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

As seen in Table 6, the Language and Spelling subcomponents have the largest gains. The tables that follow will investigate whether these gains, specifically for Language and Spelling, may also be classified according to the three clusters of learner performance. The differences in the totals for the various subcomponents must be noted: the Spelling section was out of 20 points while the language was out of 4 points and learners made the biggest gains in a question where they were required to capitalise a letter and add a full stop.

The second highest subcomponent for the Matthew Effect as seen in Table 6, was Language; although the estimate was not statistically significant. The regressions in Table 7 provide a further analysis of this subtest through clustering learners into performance categories and interacting this with the baseline score. The highest gains, in support of the Matthew Effect hypothesis, show an additional gain of 7.22 percentage points for the top 34% of learners. This estimate is statistically significant at the 99% confidence level, although we cannot be sure that this estimate is different from the estimates of the middle performance group. Overall however, there are clear incremental gains between the different categories and we

can be confident that the estimated effect size is statistically significantly different between learners performing in the bottom 33% and those performing in the top 34%.

Table 7: Regressions of language subtest with intervention and performance categories

	(1) Overall Language Score	(2) Bottom 33% Language Score	(3) Middle 33% Language Score	(4) Top 34% Language Score
Treatment	3.70***	1.30	2.69*	7.22***
Standard error	(1.1)	(1.07)	(1.36)	(2.11)
Baseline percentage score	0.69***	0.16***	0.22***	0.59***
	(0.02)	(0.05)	(0.05)	(0.03)
Observations	2466	982	717	767
R-squared	0.45	0.08	0.14	0.38

*Note *p<0.1 **p>0.05 ***p<0.01*
All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

Thus far the analysis completed points to gains being higher for learners performing in the Top 34% with the analysis of the subcomponents pointing to Spelling and Language as key areas. The analysis of these two subcomponents points to increasing gains between the different performance groups with a clear distinction between the bottom 33% and those performing above this threshold. In understanding this further, the analysis that follows will investigate if there are other factors driving the Matthew Effect such as gender, socio-economic status, or prior exposure to English.

In Table 8, regression (1) provides the estimates for the overall gains in the Spelling Subcomponent as 1.5 percentage points. The analysis also points to higher gains for the learners scoring 34% and above in the baseline test. The estimate for the middle category of performance is 1.83 percentage points while the estimated effect size for the top category is higher at 2.29 percentage points; pointing to learners in these two categories making the most additional gains in the Spelling subcomponent. While we are 95% confident that the 1.83 percentage point estimate, and 90% confident that the 2.29 percentage point estimates are statistically different from zero; it is not clear that these estimates are different from each other. Therefore overall, there is a statistically significant difference between the performance of those above and below 33% but a larger sample size would be

required to firmly distinguish between the upper performance categories in the Spelling subcomponent.

Table 8: Regressions of spelling subtest with intervention and performance categories

	(1) Overall Spelling Score	(2) Bottom 33% Spelling Score	(3) Middle 33% Spelling Score	(4) Top 34% Spelling Score
Treatment	1.5***	0.57	1.83**	2.29*
Standard error	(0.57)	(0.58)	(0.80)	(1.36)
Baseline percentage score	0.96***	0.81***	0.80***	0.86***
	(0.02)	(0.06)	(0.06)	(0.026)
Observations	2466	982	717	767
R-squared	0.77	0.25	0.38	0.67

Note *p<0.1 **p>0.05 ***p<0.01

All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

Table 9 points to a general female advantage in the Spelling subtest, estimated at 1.68 percentage points, an estimate that is statistically significant at the 95% confidence level. We can thus be sure that the impact is statistically significantly different from zero. At the same overall level, males perform somewhat worse than females at 1.20 percentage points, at the 90% confidence interval although we cannot be certain that the estimates are different from each other. To be clear, these estimates are the treatment effect for learners with gains of 1.68 for females and 1.20 for males rather than whether males perform better than females. However, females gained more, and thus have an advantage that persists at all levels except for the top 34% where males gained 2.75 percentage points, an estimate that is statistically significantly different from zero at the 90% confidence level; while females gained considerably less at 1.75 percentage points. The estimate for females is not statistically significant and we can thus not be sure that it is different from zero. Again, the issue of a reduced sample size alluded to earlier may be a contributing factor for this.

Table 9: Regressions of spelling subtest and performance categories with intervention and performance categories

	(1) Male Overall Spelling Score	(2) Bottom 33% Male	(3) Top 34% Male	(4) Female Overall Spelling Score	(5) Bottom 33% Female	(6) Top 34% Female
Treatment	1.20*	-0.13	2.75*	1.94**	1.68*	1.75
Standard error	(0.70)	(0.70)	(1.59)	(0.75)	(1.00)	(1.7)
Baseline percentage score	0.99***	0.81***	0.88***	0.94***	0.84***	0.86***
	(0.02)	(0.07)	(0.05)	(0.02)	(0.09)	(0.03)
Observations	1264	601	290	1202	381	477
R-squared	0.78	0.27	0.76	0.76	0.32	0.63

*Note *p<0.1 **p>0.05 ***p<0.01*

All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

The higher estimates of the treatment effect for females compared to males are more clearly illustrated in the Language subtest analysis in Table 10, although it must be noted that the estimates overlap and we can therefore not be sure that they are different from each other. The estimates for females are higher than those for males in every category except in the Top 34%, a pattern also seen in Table 9. The overall estimated additional effect size for females in the Language subtest is 4.38 percentage points, an estimate that is statistically significant at the 99% confidence interval while the estimate for males is much lower at 2.95 percentage points, an estimate of additional gains that is statistically significant at the 95% confidence interval. Again, caution should be taken when interpreting these results; this table shows the effect of the treatment for males and for females. While the estimated treatment effect is higher usually for females the confidence intervals overlap in every category and we thus cannot be sure that they are not the same.

Table 10: Regressions of language subtest and performance categories with intervention by gender

	(1) Male Overall Language Score	(2) Bottom 33% Male	(3) Top 34% Male	(4) Female Overall Language Score	(5) Bottom 33% Female	(6) Top 34% Female
Treatment	2.95**	1.60	8.39***	4.38***	0.28	6.3**
Standard error	(1.26)	(1.09)	(2.64)	(1.44)	(1.61)	(2.48)
Baseline percentage score	0.69***	0.24***	0.55***	0.69***	0.02	0.62***
	(0.02)	(0.67)	(0.05)	(0.03)	(0.08)	(0.04)
Observations	1264	601	290	1202	381	477
R-squared	0.43	0.10	0.43	0.46	0.16	0.39

*Note *p<0.1 **p>0.05 ***p<0.01*
All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

4.6 Conclusion

As discussed above, the main results from the study were not statistically significant; both the Intervention Group and Comparison Group made gains of similar margins. Notwithstanding this, the analysis of the main results clearly points to differentiation in performance based on gender in both the baseline and the endline. Females outperformed males by approximately 7 percentage points in the baseline with this gap being slightly larger at the endline. This could have arguably been identified in existing assessment data such as the ANAs collected by the DBE. However, limitations in the test design, administration and marking, as well as inconsistency in the data mean that there are serious constraints on the confidence and measurement of learner performance emerging form these kinds of assessments. Therefore, the conditions of testing, including the availability of a comparison group are powerful contributors from the RCT method for measuring the impact of the intervention with precision.

The second aspect of the analysis, illustrates the contribution that is possible when administrative data is interacted with learner performance data. Such analysis and research using accurate assessment information and administrative data on learner characteristics provides broader research and analysis opportunities that contribute to understanding the specific conditions under which interventions such as the RCUP make a difference. This is illustrated in the statistically significant estimate of the Matthew Effect observed in the top 34% of learners in the Spelling and Language subtests. The analysis points to a causal relationship in gains from the intervention based on prior knowledge with a clear distinction between the lowest performing 33% of learners and those above. Attempts to establish causal relationships between gender and the Matthew Relationship did not prove to be significant or distinctive but rather provided clear evidence that gains were only associated with prior knowledge and not gender, although the advantage by females from the baseline was flagged. However, males gained slightly more from the intervention for the Top 34% of learners although the estimates for this category overlap.

It is not clear from the data when the gender gap is developed as the RCUP was implemented at the Grade 4 level. Further analysis to establish the extent of the learner performance disparities by gender at the beginning of schooling would contribute towards understanding the gender correlation better. This would require the analysis of data from the Foundation Phase grades and as close to the start of schooling as possible. The EGRS RCT which evaluated a Grade 1 intervention may be useful for responding to these emerging questions. In addition, the analysis of Grade 1 baseline data collected at the start of schooling would also contribute to understanding if there are already different performance categories at the beginning of schooling. This would have implications for when remedial programmes and interventions should be considered based on a large dataset where we can be certain that the estimates of the relationship are statistically different from zero.

CHAPTER 5: CASE STUDY 2: EARLY GRADE READING STUDY

5.1 Background

The Early Grade Reading Study (EGRS) is being conducted by the Department of Basic Education (DBE) with several partners and donors and has recently (February, 2015) commenced in 230 schools in the North West province. The analysis presented here on the EGRS has benefited from direct involvement of the scholar in the research process through working at the DBE in the research and evaluation unit. The EGRS has been crafted to evaluate three competing interventions all aimed at improving home language (Setswana) reading acquisition in Grades 1 and 2. The RCT is intended to evaluate the causal impacts of three interventions: (i) a teacher training course focused specifically on the teaching of Setswana reading and literacy, accompanied by scripted lesson plans and graded reading materials; (ii) an on-site support programme to teachers from reading coaches, accompanied by scripted lesson plans and graded reading materials; and (iii) and a package designed to improve parent involvement in – and monitoring of – learning to read. Each intervention is being implemented in 50 schools within the sample. A further 80 schools have been selected as the comparison group (Department of Basic Education, 2014). As was the case in the Pinetown RCUP study, separate organizations have been contracted to undertake the implementation of interventions and the data collection for evaluation, with the evaluation agent being blind to which group schools are part of.

The interventions were intended for the full Grade 1 academic year in 2015 and Grade 2 in 2016; however classroom level implementation in 2015 took place from Term 2 as Term 1 was used for the training of teachers for Intervention 1 and 2, as well as the recruiting and training of Community Reading Coaches (CRCs) for Intervention 3.

Interventions 1 and 2 consist of scripted Setswana lesson plans for Grade 1 (2015) and Grade 2 (2016) as well as 16 titles of Vula Bula Setswana Graded Readers per grade; 3 posters; flashcards of ‘look and say’ words as well as phonic sounds and words; pattern booklets for Grade 1 handwriting; and Knowledge Tree DVDs on issues including classroom management, learning environment promotion, management of resources, and language

methodologies. The distinguishing factor between the two interventions is the model used for teacher training. For Intervention 1 teachers are trained at the beginning of Terms 1 and 3 for two days at a centralised venue. In Intervention 2 teachers receive training in smaller clusters through reading coaches at the beginning of each term as well as a monthly visit by a reading coach to provide classroom-based support. Intervention 3 is focused on parental involvement delivered by a CRC for the parents of Grade 1 learners in 2015 and of Grade 2 learners in 2016. The content of the material focuses on understanding how children learn to read, games and activities to build pre-reading skills, aural and visual comprehension skills, and development of homework programmes. In addition the Platinum Series Le Re Tlhabetse Readers have been handed to parents (Department of Basic Education, 2015).

The Theory of Change (TOC) for the interventions is follows literature on how reading acquisition takes place; with the main path being that the possession of the appropriate vocabulary and the ability to decode when practiced lead to fluency and ultimately comprehension. Exposure to language within society provides learners with a language-specific vocabulary as well as the beginning of the development of phonological awareness- the ability to hear sounds and segmentation within words. However, the next steps in reading, namely decoding and mastery of phonemic awareness, require systematic teaching and practice to develop into reading acquisition.

In terms of the materials and lesson organisation the TOC for Intervention 1 and 2 is based on this; providing coherent, well-paced, sequenced lesson plans that have supporting resources to further develop vocabulary and phonemic awareness, as well as opportunities to practice. The support provided through either the reading coach or the centralised training are alternative mechanisms to build this repertoire - through both capacity building and local accountability - amongst teachers who are the instructional custodians of the delivery of the programmes.

Intervention 3 on the other hand, has a different TOC but the same outcome. Much has been written about the role of parents and the difference social capital can make within the schooling context. This is largely based on developed countries but there is a growing body of knowledge from developing countries. Little is known about the possibilities and the

impact of parental involvement in the South African context. This is particularly true for poorer learners within the system. The rationale for this intervention then, is to begin to establish an evidence-base and explore the possible impact of parental involvement on reading acquisition, being cognisant of the low education levels amongst these parents.

5.2 Sampling and Data Collection

The sample of 230 schools consists of non-fee paying schools in the Dr Kenneth Kaunda and Ngaka Modiri Molema districts of the North West province. The sample was restricted to ordinary public schools that are classified as Quintiles 1 to 3 according to National Treasury poverty classifications and offer Setswana as the Home Language in the Foundation Phase. Multi-grade schools, boarding schools and schools with less than 20 or more than 120 learners were excluded (Department of Basic Education, 2014).

The sample was stratified into 10 strata of 23 similar schools, based on school size, SES and previous performance in the ANAs, prior to randomisation which was done through a computerised lottery. Based on the sample size and the testing of 20 Grade 1 learners per school in 2015 and the same 20 Grade 2 learners in 2016, the study is powered to detect a minimum effect size of 0.21 standard deviations when comparing an intervention group (50 schools) with the comparison group (80 schools). When comparing between treatment groups (each of 50 schools) the minimum detectable effect size is 0.23 standard deviations. Additional main assumptions are a 95% confidence interval, power of 0.8, an intra-class correlation coefficient (rho) of 0.3 and a correlation between the pre- and post-test of 0.7 (Department on Basic Education, 2015).

5.2.1 Baseline Data Collection

Baseline data collection commenced on 4 February 2015 and concluded on 24 February 2015 in all 230 participating schools. The midline data collection at the end of Grade 1 was then completed in October and November 2015. However, the analysis in this study will be limited to the baseline data collected as the midline data had not been available for analysis during the duration of this study.

Fieldworkers were expected to test a random sample of 20 learners per school using a specific sampling procedure they were trained to administer. Based on the administration process and the fact that the assessment agency was blind to which schools were allocated to the various research groups, there is no reason to expect bias in data collection. The realised sample from the baseline assessment was 4539 learners out of an expected 4600 learners. In 204 schools, exactly 20 learners were tested. In 4 schools there were 21 learners tested, and in 12 schools 19 learners were tested, although it is not clear why this occurred. The remaining 10 schools had between 9 and 17 learners tested.

Detailed information on the sampling process, the data collection, the intervention and assessment components are available in the main report[9] on the baseline data collection. A letter of permission to use the EGRS data is attached as Appendix F.

5.3 Learner Characteristics

This section provides some descriptive information on learner characteristics across the entire sample of 230 schools. The data analysis completed in this section is mostly based on replicating descriptive data from the official baseline report.

Table 11: Learner characteristics by gender and age

Learner Age	No. of Female Learners	Cumulative % of Female Learners	No. of Male Learners	Cumulative % of Male Learners	Total No. of Learners	Cumulative % of all Learners
5	227	14.23%	230	12.41%	457	13.25%
6	768	62.38%	862	58.90	1,630	60.51%
7	483	92.66%	577	90.02	1,060	91.24%
8	94	98.55%	145	97.84	239	98.17%
9	17	99.62%	31	99.51	48	99.56%
10	6	100%	9	100%	15	100%
Total	**1595**	**100%**	**1854**	**100%**	**3,449**	100%

The age groups indicated in the table above exclude data for learners that reported to be younger than 4 or older than 10.

The modal average age in the EGRS sample is 6 years old with approximately 91% of learners aged 7 and below, which is in accordance with the enrolment policies for the start

[9] The baseline report for the EGRS is attached as Appendix E

of school. Similar to the RCUP sample, the males in the sample are older than the females. Approximately 62% of females are 6 years or younger while about 59% of males fall into the same category.

The proportion of males in Grade 1 is substantially higher than the proportion of females; this may point to the manifestation of the lags in performance and progression seen in the RCUP starting from the Grade 1 level. Males may be repeating Grade 1 more frequently than females or starting Grade 1 later due to delayed progression at the Grade R level. Literature on system efficiency does point to negative correlations between age and learners' performance in the majority of schools, as repetition is not clearly aligned to appropriate remediation (Department of Basic Education, 2014).

Table 12: Learner characteristics by location and school quintile

District	Quintile			Total Number of Learners
	1	2	3	
Ngaka Modiri Molema	1,676	1,109	698	3,483
Dr Kenneth Kaunda	522	180	354	1,056
Total	**2,198**	**1,289**	**1,052**	**4,539**

Source: Replication of analysis from EGRS data.

The Dr Kenneth Kaunda and Ngaka Modiri Molema districts are part of the four education districts in the North West province. The main language spoken across the province is Setswana, spoken by 42% of the population, followed by Afrikaans at 22% and English at 20% (DBE, 2015). The province is relatively poor, with no-fee paying schools constituting 81% of schools in Dr Kenneth Kaunda and 91% in Ngaka Modiri Molema.

According to the DBE District Profiles (2015), there are 164 634 learners and 248 schools in the Dr Kenneth Kaunda District. Primary schools are the biggest schooling phase at 60%, consisting of 149 schools, although 1% of school are intermediate schools and 17% are combined schools. The majority of schools are long distances away from the district offices, 25% of schools are between 25 and 50 kms away; and 55% of schools are greater than 50 kms away.

The Ngaka Modiri Molema district is slightly larger in terms of learner numbers, serving 188 429 learners through a total of 404 schools. Primary schools are similarly the largest phase, consisting of 247 schools or 61% overall. There are a similar proportion of combined schools, 17%, with 3% being Intermediate schools. Distances between the district offices and schools are similarly large: 23% are between 25 and 30 kms, and 47% greater than 50 kms away (DBE, 2015).

Table 12 shows that the largest number of learners (approximately 48%) participating in the study, are from Quintile 1 schools. The highest numbers of these learners are in the Ngaka Modiri Molema District which has almost triple the number of Quintile 1 learners as Dr Kenneth Kaunda District.

5.4 Learner Performance

The main instrument administered was a Setswana learner test mostly based on the EGRA in terms of letter recognition fluency, word recognition fluency and sentence reading components. In addition, the test contained a picture comprehension test or expressive vocabulary test, a phonemic awareness component, and a digit span memory component. The test focused on pre-literacy skills as well as competencies that are strong predictors of reading according to education theory (Gove and Wetterberg, 2011). The underlying construct that the baseline test attempted to measure is reading and pre-reading ability as well as skills known to be predictive of learning to read, such as the child's working memory The learner test was orally administered by the fieldworkers to each individual learner, lasting no longer than 20 minutes with opt-out rules incorporated for items learners could not respond to. In addition a teacher questionnaire with a short teacher reading fluency test, a home background questionnaire, and a school principal questionnaire formed part of the instruments.

As discussed in Chapter Two, the two main baseline contributions are a reduction in the variability when interpreting the endline data, as performance in the baseline is predictive of performance in the endline. Secondly, the baseline confirms balance across the

intervention and comparison groups based on observables. The baseline measured learner performance prior to the implementation of any interventions.

Table 13 provides a summary of learner performance across all subcomponents of the baseline test. Learners scored the highest on the picture comprehension (Section A) which was intentionally designed to be the easiest aspect of the test, with most learners scoring an average of 8.58 out of 10 marks. The pictures consisted of common Setswana actions and nouns, and efforts were made to exclude pictures with commonly known borrowed words from other languages. The lowest scores were in the harder component of the test, which were included from the EGRA to measure learner competencies expected from learners with substantive prior exposure to curriculum content and who could read. It must be noted that low learner scores in these areas do not warrant concern as reading ability is a competency learners are only expected to mastered in Grade 1.

The first of these components of low-performance were the words correct subtest, with the average learner score of 1.91 marks although a few learners scored 50 marks. The second was the words correct in a sentences subtest, where learners scored an average of 1.22 marks out of 15 marks although as expected, a few learners scored full marks. Most learners could only read the simpler words within the first sentence but not the entire sentence. The third subtest was the sentence reading comprehension out of 3 marks, which was based on the sentences read in the previous section, scoring an average of 0.73 marks.

Table 13: Summary statistics on baseline learner performance

Baseline Test Subcomponent	Observations	Mean	Std. Dev.	Min	Max
Picture comprehension /10	5439	8.58	1.4	0	10
Letters correct	4452	5.08	9.86	0	99
Digit span words /5	4539	2.45	1.31	0	5
Digit span numbers /5	4539	2.54	1.34	0	5
Digit span total /10	4539	4.99	2.43	0	10
Phonemic awareness /12	4539	2.17	3.11	0	12
Words correct	4447	1.91	5.28	0	50
Number of sentences word correct /15	4539	1.22	3.38	0	15
Sentence reading comprehension /3	4539	0.73	1.24	0	3
Combined score mean 0 SD 1 pca	4385	0	1	-2	5

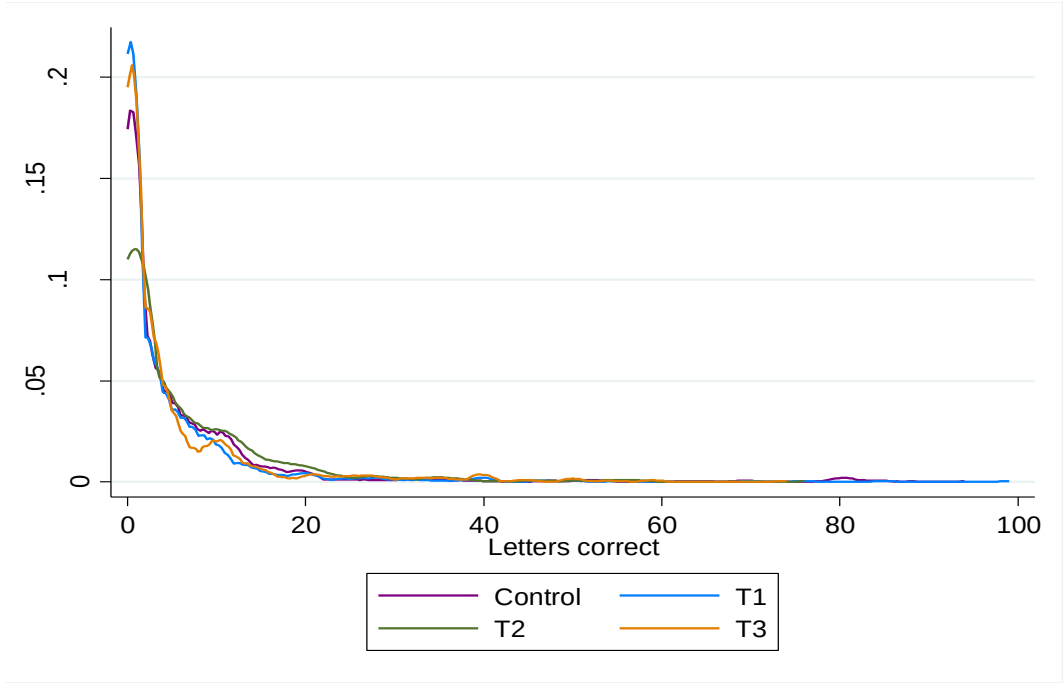
Source: Replication of analysis from EGRS data.

Although both floor and ceiling effects were detected in the various subcomponents of the test, when the different measures were combined using Principal Component Analysis (PCA) the measure gave a normal distribution of learning outcomes with neither a floor effect nor a ceiling effect. The implications of this are that changes in the learning outcomes across the distribution should be detectable and as such the baseline test will be useful in correlating impact with endline performance. The PCA calculation, which is not shown in this analysis, is particularly useful in equating different aspects of tests containing various embedded steps which made traditional methods of calculating a combined score difficult.

This method is therefore useful in calculating the EGRS baseline score based on the nature of the test. For example, the letter recognition subtest had over 30 alphabets that learners had to identify within a minute. Scoring had two components, letters reached, and letters correct. While there may be more items and somewhat complex scoring within this subtest, this component of the test was not the most difficult. The sentence reading subtest was shorter than the letter recognition subtest in terms of items and simpler to score but more cognitively demanding. In order to meaningfully equate the various components, the PCA calculation assigns weights to each subtest based on the assumption that the primary linear combination which captures the most common variation amongst the variables included, represents the underlying construct of interest. In this case we might think of the primary underlying construct being measured as reading ability or pre-reading ability. The weight given to each variable when calculating the total composite score is then determined by the extent of that variable's correlation with the first principal component. The intuition is that a subtest that is not well correlated with the other subtest may be measuring something different from the intended underlying construct – in that case, that subtest should therefore carry less weight in a composite index (Department of Basic Education, 2015).

The performance of learners in two subtests is shown in the section that follows. These figures show that are no systematic differences in performance across the different interventions, and we can confidently attribute differences that may occur in learner performance measured in the midline or endline tests to the impact of the interventions.

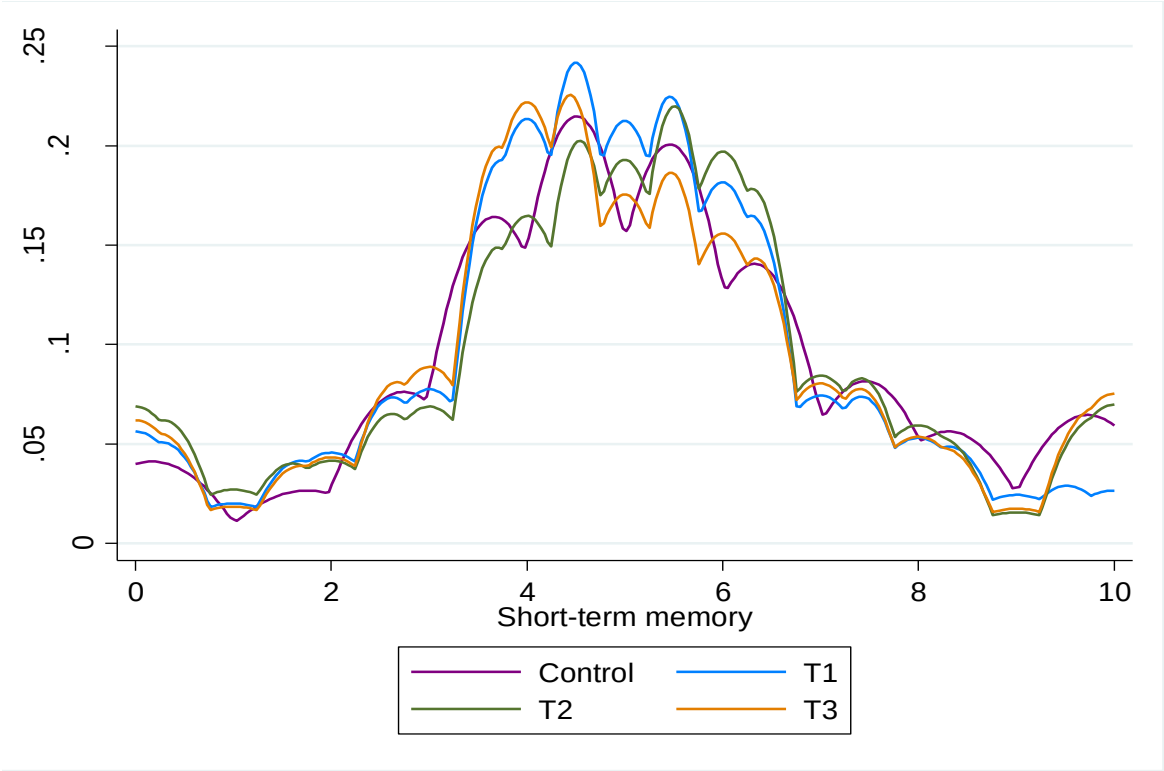
Figure 7: Kernel density curves for Section B (letters correct) by treatment arm



Source: Replication of analysis from EGRS data.

Figure 7 shows the distribution of learner performance for section B of the test on letters correct. The subtest entailed the fieldworker pointing to each word and asking learners to say the letter sound within 60 seconds. Approximately 42% of learners could pronounce any letter sounds. Learner performance across the four groups was near identical, and as such affirms the success in randomisation resulting in balanced treatment assignment.

Figure 8: Kernel density curves for Section C (Short-term memory) by treatment arm



Source: Replication of analysis from EGRS data.

Figure 8 shows the distribution of learner performance on the digit span subtest comprising of both the word and number digit span memory subtests. In the numbers component, the fieldworker read out loud two unrelated numbers in Setswana to a learner and asked the learner to repeat these back in the same order. This was increased to three numbers, then four. The same pattern was followed for the word component. As indicated in Table 13, leaners scored an average of 50% in this entire section.

As noted in the RCUP data there were clear differences in performance within the same sample of learners with a proportion of learners performing far above their peers. The learner performance gaps were associated with specific characteristics including gender and prior knowledge. A brief preliminary analysis was similarly done with the EGRS data as far as possible considering that the data analysis was limited to baseline data. The focus was to determining whether inequalities in learner performance exist at the very start of formal schooling and if there are any gender advantages.

Table 14: Learner performance between the 10th and 90th percentile – letters correct and digit span

Percentiles of the baseline test	Percentiles of learner performance - letters correct	Percentiles of learner performance - digit span
10%	0	2
25%	2	3
50%	6.5	4
75%	12.5	5
90%	24	6
Mean	9.96	3.95
Standard Deviation	12.99	1.91
No of observations	4539	4539
Range in learner performance	(-4, 80.5)	(0.10)

Source: Own calculations from EGRS data.

Table 14 shows the range in learner performance between the 90th and 10th percentile of learners across the sample of schools in the letters correct subtest and the digit span subtest respectively. The table arguably shows that there is a wide ability and cognitive development range amongst learners across the schools in the sample, a reality that was highlighted across the various international RCTs in Chapter Three with regards to learner performance. These differences are most clear in the letters correct subtest, learners performing in the 25th percentile scored 2 while learners in the 75th percentile scored 12.5 and learners in the 90th percentile scored 24. This relatively broad range, even between the 25th and 50th percentile reflects the difference in learner readiness to learn at the start of school. The inequality across learners provides a difficult learning environment and points to these differences being evident from the start of schooling. Noting that short-term memory, tested in the digit span subtest is highly predictive of future learning ability, the difference of 3 marks between the 25th percentile (3) and the 90th percentile (6) in the digit span subtest is particularly concerning as they relate to predicted future reading ability and not only attained knowledge at the start of school as in the letters correct subtest.

Table 15: Regression of baseline score controlling for parental writing ability and gender

Variable	Average score (1)	Letters correct (2)	Digital Span Memory Total (3)
Female Learner	0.11**	1.03**	0.16**
	(0.32)	(0.29)	(0.8)
Parent writes	0.02***	0.55*	-0.05
	(0.37)	(0.33)	(0.09)
Learner age	0.01	1.27**	-0.02
	(0.02)	(-2.09)	(0.06)
Constant	0.21	-4.2	5.8
	-0.23	-1.44	(0.38)
Observations	3447	3498	3575
R-squared	0.1	0.04	0.08

Source: Replication of analysis from EGRS data.

*Note *p<0.1 **p>0.05 ***p<0.01*
All models include controls stratification dummies, learner gender, learner age. Standard errors are adjusted for the fact that learners are clustered in schools.

The analysis of the RCUP data showed a difference in performance by gender favouring females, in addition to higher enrolment numbers and older learners amongst males. The analysis of the learner characteristics in Table 11 clearly showed similar patterns of enrolment and age. Table 15 also showed a female advantage in learner performance in the EGRS baseline, this time at the very start of grade 1. The estimates for the average performance: 0.11 letters correct; 1.03 and digit span; 0.16 are all positive and significant at the 95% level of confidence. The table also shows the effect of parental education and gender on the baseline score. Learners with parents that could write scored an additional 0.02 percentage points on average, an estimate that is statistically significant at the 99% level of confidence. The estimates for the letters correct also showed an additional gain of 0.55 percentage points which is significant at the 90% level of confidence. Although the estimate for the digit span subtest is negative, it overlaps with zero and is therefore not statistically significant.

5.5 Conclusion

The analysis of the EGRS data shows the same learner characteristics patterns as the RCUP data analysis; pointing to even larger differences in learner enrolment by gender, as well as older males than females. This trend is clearly not established later in the schooling system but is evident at the onset of schooling as seen in the EGRS sample. Further analysis of learner baseline performance similarly points to a female learner performance advantage at the start of schooling. This is particularly stark as learners were tested in their home language, a language we intuitively expect learners to be equally equipped to learn in. The reasons for this are unclear, however, education literature points to males being weaker readers than females.

The wide range in learner performance at between the 10^{th} and 90^{th} percentile highlight the challenges the education system, specifically teachers, are faced with in determining how to translate and deliver the curriculum considering the broad range of ability and knowledge between learners at the start of schooling. This supports the assertions made in Chapter Three for improving learner performance, namely, the acknowledgement that there is a wide range of learner ability within similar schools and a diverse approach to teaching through ability groups or additional differentiated targeted support are key in improving learner performance.

The interaction of learner performance with parental education also clearly shows a statistically significant advantage for those learners that have educated parents. This may explain why specific learners start school at a higher performance level. Literature says that parents with higher educational attainment value education and thus make more efforts in literacy efforts and supervision, which positively affect learner performance. The most concerning aspects of this trajectory are that it points to the urgent need for early remedial learning, even at the Grade 1 level to ensure that the inequality in learner performance is reduced and a larger proportion of learners remain on par with the curriculum expectations.

What seems to be emerging is that the inequality in learner performance is present at the beginning of schooling, persists and widens in the Foundation Phase as seen in the RCUP.

The variance in learner performance at the beginning of schooling provides a reasonable argument for why the Matthew Effect is evident in later grades. It is also clear that there is a broad range of performance within the same classrooms and across schools.

The EGRS may be seen as a more extensive follow-up to the RCUP to answer some of the unanswered. Some of the answers have begun to emerge as explored through the baseline data but remaining questions that would be answered as the RCT continues include whether an intervention that is implemented over a longer duration have an impact? Can intervening right at the start of school be a strategic point to intervene? And Can a HL literacy intervention have lasting educational benefits?

CHAPTER 6: CONCLUSION

This study began by describing the education landscape in South Africa with the aim of substantiating that there is a legitimate concern about the quality of learning and teaching in the education system. The low achievement levels of South African learners in local and international assessments in Mathematics and Literacy has been demonstrated extensively in several reports and academic publication, some of which have been referred to in the study. The assessments referred to include the ANAs, TIMSS, SACMEQ, PIRLS and pre-PIRLS. Specifically in language, the PIRLS 2006 study showed that more than 80% of South African children had not yet learned to read with meaning by grade 5. The pre-PIRLS results of 2011 indicated that about 29% of Grade 4 learners did not have the fundamental reading skills required at a Grade 2 level. Learners tested in African languages, particularly Sepedi and Tshivenda, were especially likely to have not reached this basic minimum level of reading (University of Pretoria, 2012).

Underneath the comparatively weak national average performance of South African learners, there is a learning inequality gap that especially disadvantages poorer learners. Through an analysis of PIRLS and pre-PIRLS, evidence indicates that learners in urban areas are approximately two years' worth of learning ahead of those in rural or township areas (University of Pretoria, 2012). This is a pattern seen across the various national assessments administered. The patterns in performance affirm the assertion that the education system is bimodal- there are two distinct modes or peaks in the distribution of learner performance. This pattern of inequality remains aligned to historical school endowment with urban, former White and Indian schools retaining an advantage. This is problematic in a range of spheres- politically, socially and economically. The process of converting the underperforming schools into high performance schools is clearly necessary for a range of reasons including but not limited to education being the most sustainable and long-reaching national effort to bring about equality in outcomes in South Africa across different SES group and different races. However, as pointed out in Chapter 1, meaningful systemic change requires an extended process based on evidence. There is rich literature that provides details on the challenges and reasons this change is necessary but there is a limited

resource base of rigorous research and evaluations that may be implemented at scale in response.

6.1 Summary of Literature: Systematic Literature Review on the RCT Methodology

Chapter Two provided a literature-based argument for the use of RCTs in education research by providing a methodological argument for the strengths, appropriateness and applications of the method in education research. Several quantitative internally valid methods were discussed, however this argument was largely based on the nature of the RCT design which lends itself to being the cleanest method of measuring impact and thus potentially making a substantial empirical contribution to policy-level intervention options. As discussed in Chapter Two, this is achieved mainly through randomisation. By using a lottery to allocate participants to an intervention and a control group, RCTs constructs a credible 'counterfactual' scenario – what might have happened to those who received an intervention had they not received it. This addresses selection bias as recipients have an equal chance of being allocated to be participants or non-participants and thus there is no reason to expect systematic differences between intervention recipients and non-recipients and when measuring outcomes we can be certain that we are measuring the true impact of the intervention or policy rather than other confounding factors. Furthermore, the large sample size required in RCTs as part of the technical parameters may be selected to reflect the larger schooling population and thus findings from an RCT may arguably be representative of a larger population and applicable in the same way. This scientifically rigorous method allows researchers to apply econometric methods that test the statistically validity of the estimates of effects, which would provide precision about the impact measured from the interventions implemented.

RCTs, similar to all other methodologies, are not without their limitations. There are various critiques including epistemological objections, lack of clarity on the methodology, and technical methodological concerns about how RCTs are designed, implemented and interpreted. The main critique is external validity – the extent to which the RCT represents a broader population, in other words, the extent to which findings from an RCT may be interpreted beyond the sample in the study. A robust discussion of the critiques has been

presented in Chapter Two with steps that may be taken to mitigate these discussed. One of the main arguments made in this study is that the limitations of RCTs do not outweigh the value gained in research and educational terms, from using this method. RCTs should be a methodological option for education researcher and policy-makers, but not the only option, as they are not appropriate to respond to all policy questions.

6.2 Summary of Literature: Literature Review on International and National Experimental and Quasi-experimental Research in Education

As discussed in Chapter Three, there is a large body of knowledge on international experimental and quasi-experimental research focusing on access and participation. The most effective interventions in this area include monetary incentives through conditional grants, cash transfers or reduced fees to mention a few; as well as school-based interventions that include improving learner wellbeing and health. The second focus based on international literature, was learner performance, and although there are a substantial number of studies with some recommendations on the most efficient interventions, these did not emerge as clearly as in the access and participation theme. Nonetheless, the literature firstly highlights that resources as a singular input, have limited impact on improving learning. This seems to be the case in South Africa, there have been initial gains from the provision of textbooks and other LTSMs based on assessments and other evaluations, however the gains plateau in the long run. The most effective interventions are centred on the provision of well-constructed curriculum content, mediated by specific, targeted teacher training and supported by high quality LTSMs. This may then be complimented by accountability for maximum effectiveness. Without providing these various resources in combination, increased accountability as in the Pratham study in India or activity did not result in sustained effects on learner performance. The second aspect that emerged from the international literature was based on accepting that there are learning inequalities within schools and responding to these through delivering differentiated support or learning mediated through smaller, ability-based classes or using computer-based technology to support learning at the various paces of learners.

The details of these interventions are discussed in the chapter. The main emphasis between the category of participation and access; and learner performance is that interventions targeted at access and participation hardly include a curriculum component and this may be a gap worth addressing; the interventions targeting access and participation hardly ever seem to impact on learning outcomes with the exception of initial gains from increased exposure to schooling where access or attendance were previously low; and even in the most prominent health-based intervention, which was the deworming in Kenya as well as in seven other studies, the effect on learning outcomes was close to zero. The reality of education inequalities within schools and classrooms is a fact that is known but that has not started to form part of policy in the education sector, especially in South Africa. However the literature points to this as an international phenomenon that should be the starting premise of interventions to improve learning outcomes once the basic resources and general improvement aspects have been addressed. This is an area that requires serious consideration for policy makers and is relevant for South Africa. It also points to the urgency required to research and implement interventions that begin to narrow this gap.

The section of the chapter that focused on literature on reading interventions points to the same issues as those raised in the learner performance literature, however, the issue of inequality clearly begins to highlight a marked effect of intervention on learners based on their prior knowledge. In some of the studies reviewed, there was a clear Matthew Effect - the notion that initially better-performing children typically gain more from additional interventions and from schooling itself. This seemed to result from the reality of incongruence in the knowledge levels of most learners within schools and the curriculum or cognitive demand of the resources provided, as in the Kenya textbook study. The study pointed to the need to acknowledge these differences and remediate these if interventions are to have an effect for the majority of learners. The literature also reflected that the learner ability level that programmes are target at, as well as the institutional infrastructure supporting their delivery, make a difference. Some of the studies pointed to larger gains for previous non- readers, noting that they were the intended recipients in the development of the material but there were still differential gains based on the prior ability of learners. This is encouraging as it shows that even learners at the bottom end of the performance spectrum within schools may be remediated effectively when programmes are designed

with these learners in mind. The success of interventions when they were more integrated into the schooling system as additional resources or support to normal practices, as well as a higher rate of success when interventions were implemented over a longer period are important and educationally meaningful findings. They reflect that when school management and administration are implemented correctly, they make a contribution to effective learning. These insights are directly relevant for the South African education context and warrant consideration as a contribution as a point of reference for the design of future interventions.

The limited number of credible literature that observes the methodological prescripts of experimental and quasi-experimental methods nationally, supports the claim made in this study: that the use of RCT's which are but one methodological option, is not optimised by education researchers or policy makers. In as much as RCTs are not always appropriate to respond to all policy and education questions, there are close to none specifically in reading and this cannot be due to their inappropriateness but rather it is due to the limitations in the methodological options considered for education research and policy development.

6.3 Summary of the Case Studies: RCUP and EGRS

The main finding from the RCUP RCT was that although learners in intervention schools improved their test scores between the baseline and the endline assessment, the learners in comparison schools improved by a similar margin. Overall gains were not statistically significantly different between the treatment and control groups. Even though the findings did not affirm the hypothesis that the RCUP was an effective intervention that should be used scaled-up nationally it yielded several important education intervention lessons and raises several questions for further research. Firstly, it is evident that seemingly well-designed programmes may not have as large an effect as one might expect in the absence of a rigorous evaluation. This illustrated the importance of using internally valid research methods for evaluating programme impact. This case study should also make it clear that the importance of a valid estimate of the counterfactual cannot be overstated. The study by Hellman which showed large gains had fundamental methodological limitations and thus probably concluded a false positive or an exaggerated estimate of the impact. The

prevalence of such practices by education researchers and the use of such evidence by policymakers warrants serious concern.

The finding also point out that apart from the design quality of a programme and the integrity of implementation, there may be contextual factors pertaining to learners, schools and communities which either preclude or are conducive to the effectiveness of an intervention. The findings from the analysis made possible by the depth of the RCT data, have a secondary, equally important contribution to education policy through an empirical investigation, namely, that one of the key premises in education research is realising that programmes have a differential impact on learners depending on their prior knowledge. The Matthew Effect emerging from the analysis supports this assertion for the South African context and has implications for future remedial programmes. The methodological advantages of the RCT; namely a large sample size that is reflecting of the larger schooling population, and the ability to provide statistically valid estimates of effects, make the insights gained from this analysis meaningful for the larger education sector.

In both of the RCTs males were older than their female counterparts within the same grades and there were more males than females. In the RCUP the modal age was 9 years old, with approximately 82% of girls aged 10 or younger while this estimate was approximately only 70% for males. In the EGRS data, approximately 62% of females were 6 or younger while the estimate for males was only approximately 59%. It is clear from both data sets that there is higher efficiency in schooling for girls and this pattern persists from the start of schooling through to the start of the Intermediate Phase. The reasons for this are not clear especially at the start of schooling, although the analysis of the learner performance information provides some correlations and point to the contribution of learner background towards school readiness.

Research and learner assessment data in South Africa have continuously shown that learner performance is weak. The analysis in this study further affirms this. In the RCUP the majority of learners scored less than 20%, pointing to the early development of learning gaps. An analysis of the EGRS data points to these gaps existing at the beginning of schooling, with the RCUP data pointing to these gaps widening over time rather than narrowing.

In terms of a comparison between males and females, there seems to be a Matthew Effect for each gender respectively but the patterns are unclear when males are compared with females. However, the effect is based on prior knowledge rather than gender. Slight indications of female advantage in performance is correlated with learner performance up to 66% based on the baseline test while males seem to perform better for the top 34% of learners. The estimates for the top 34% by gender overlap, and we can thus not be certain of differences in performance by gender. Notwithstanding the Matthew Effect, females have a learning advantage in both case studies. Females performed better in the baseline, maintained these gains and started off on a higher base in most of the performance categories or subtests used to analyse the data.

The findings from the RCUP point to further research questions, some of which could be addressed by the EGRS. These include what the characteristics of learners are at the start of schooling, what the range and inequalities across learners in learner cognitive ability and knowledge are at the beginning of schooling, what learner knowledge levels exist at the beginning of schooling and whether longer interventions would make a lager difference. The analysis of the EGRS baseline data in this study provided an opportunity to respond to some of these questions. The main findings of the analysis show that inequalities are present at the start of schooling; there is a broad range in learner performance already established in grade 1. The analysis also showed higher performance by females and the interaction of learner performance with parental education also clearly showed a statistically significant advantage for those learners that have educated parents.

6.4 Concluding Remarks

This study has clearly responded to the two main research questions: what is the value of RCTs as a methodological option for education research geared towards informing policy? and How do case studies of RCTs illustrate the value of the methodology in gaining insights for policy and practice? There is a case to be made for the use of RCTs as an important methodology for policy makers and researchers in education based on the internal validity

of the method, its relevance to the larger population, and rich analysis that may continue to be done with RCT data.

The study also contributes to the knowledge and practices in addressing the core of the education challenges that persist. The state has a mandate to use policy to attempt to address challenges in society either through preventative measures, regulation of specific practices or introducing punitive measures for behaviour viewed as contravening the Constitution of South Africa. How policy makers respond to the status quo in education affects the broader society; shaping and informing public activities, informing norms and in the long term shaping society and history. The framework of national and provincial policies that support institutional and resource-based change in the system is firmly established; and there is a rich body of diagnostic research on system deficiencies; but the third critical component, empirical evidence-based South African research in support of policy interventions specifically targeted at improving learner performance for the majority of learners, is lagging behind.

The policy making process is based on various factors and inputs which includes research and evaluations. There is a distinct role that academics can play as individuals or collectively through higher education institutions, in informing the education policy making process but there is currently a clear gap in education research on rigorous evidence on classroom-based practice and resources that have a measurable effect on learner performance in a developing country like South Africa. Research that evaluates interventions is often based on case studies or small pilot studies. As valuable as these methods are in helping us to understand classroom and school dynamics and identify what makes interventions work, they are often resource-intensive and difficult to replicate or scale up. They are useful in their small-scale detail, but larger samples that will allow us to generalise to a larger population are required to inform policy decisions. The policy process is impoverished as there is a lack of research meeting the following two criteria: the development or evaluation of interventions and findings that are relevant to the larger school population; and the establishment of causal impact of interventions and policies measured precisely.

7. REFERENCES

Aiken, A.M., Davey, C., Hargreaves, J.R. & Hayes, R.J. (2014). Reanalysis of health and educational impacts of a school-based deworming programme in western Kenya: Part 1, pure replication. 3ie replication paper 3. Part 1. Washington, DC.

Abeberese,A., Kumler,T. & Linden,L. (2011). Improving Reading Skills by Encouraging Children to Read: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines. Working Paper: Columbia University.

Angrist,J. & Krueger, A. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 15, 4: 69-85.

Böhmer, B. (2014). Testing Numeric: Evidence from a randomized controlled trial of a computer based mathematics intervention in Cape Town High Schools. Masters Thesis, University of Cape Town.

Boiseiere, M. (2004). *Determinants of Primary Education Outcomes in Developing Countries*. Washington, DC: World Bank.

Carnoy, M., Chisholm, L., Addy, N., Arends, F., Baloyi, H., Irving, M., Raab, E. et al. (2011). The process of learning in South Africa: the quality of mathematics teaching in North West Province. Technical Report. Stanford University School of Education and HSRC.

Clay, R.A. (2010). More Than One Way to Measure. American Psychological Association. www.apa.org/monitor/2010/09/trials.aspx. Accessed 12 December 2015.

Clearinghouse of Male Circumcision for HIV Prevention. Random Control Trials. https://www.malecircumcision.org/research/randomised-controlled-trials. Accessed 12 February 2016.

Combrinck, C., Van Staden, S. & Roux, K. (2014). Developing early readers: Patterns in introducing critical reading skills and strategies to South African children. *Reading & Writing* 5(1), 9 pages. http://dx.doi.org/10.4102/rw.v5i1.45.

Damon, A., Glewwe, P., Wisniewski, S. & Sun, B. (2015). What Education Policies and Programs Affect Learning and Time in School in Developing Countries? A Review of Academic and Aid Agency Evaluations from 1990 to 2014 for the Swedish Expert Group for Aid Studies (EBA).

Deaton, A.S. (2009). Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. National Bureau of Economic Research. Working Paper 14690. http://www.nber.org/papers/w14690. Accessed 26 November 2015.

Department of Basic Education. (2010). National School Nutrition Annual Report. Pretoria: Department of Basic Education.

Department of Basic Education. (2012). *Summary Report: Evidence Based Policy Review Workshop*. Pretoria: Department of Basic Education.

Department of Basic Education. (2013a). Technical Report. *Developing and Evaluating the First Phase of the Grade 12 Mind the Gap Study Guide Series*. Pretoria: Department of Basic Education.

Department of Basic Education (2013b). Macro Indicator Report. Pretoria: Department of Basic Education.

Department of Basic Education. (2014a). Annual Report 2013/2014. Pretoria: Department of Basic Education. Retrieved 15 May 2015 from www.education.gov.za.

Department of Basic Education. (2014b). Terms of Reference. Implementation of Reading Interventions for the research project: "Early Grade Reading Study". Pretoria: Department of Basic Education.

Department of Planning, Monitoring and Evaluation. (2014). *Concept for the National Evaluation Plan 2015/16.* Pretoria: Department of Planning, Monitoring and Evaluation.

Department of Planning, Monitoring and Evaluation. (2015a). *DPME Evaluation Guideline 2.2.16 Departmental Evaluation Plan.* Pretoria: Department of Planning, Monitoring and Evaluation.

Department on Basic Education. (2015b).Baseline Report : "Early Grade Reading Study".

Donaldson,S & Christie,C .(2004). The 2004 Claremont Debate: Lipsey vs. Scriven. Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard? http://evaluation.wmich.edu/jmde/

Duflo, E., Glennerster, R. & Kremer, M. (2006). Using Randomization in Development Economics Research: A Toolkit.

Edwards, S., Lilford, R., Braunholtz, D., Jackson, J., Hewison, J. & Thornton, J. (1998). Ethical issues in the design and conduct of randomised controlled trials. *Health Technology Assessment*, 2,15. http://www.journalslibrary.nihr.ac.uk/ data/assets/pdf file/0004/64822/FullReport-hta2150.pdf. Accessed 16 February 2016.

Fleisch, B. (2008). *Primary Education in Crisis: Why South African schoolchildren underachieve in reading and mathematics*. Cape Town: Juta.

Fleisch, B., Taylor, S., Schöer, V. & Mabogoane, T. (2015a). The value of randomised control trials in system-wide improvement: The case of the Reading Catch-Up Programme impact evaluation. Unpublished paper.

Fleisch, B., Taylor, S., Schoer, V. & Mabogoane,T. (2015b). Assessing the impact of the RCUP: A report of the findings of the impact evaluation of the Reading Catch-Up Programme. Unpublished paper.

Gertler, P.J., Martinez, S., Premand, P., Rawlings, L. & Vermeersch, C. (2010). *Impact Evaluations in Practice.* World Bank. Retrieved 10 October 2014 from http://www.worldbank.org/ieinpractice.

Girdwood, E. (2013). Exploring the Evidence on Early Learning: A Literature Review. Impact Evaluation of the Introduction of Grade R on Learning Outcomes. Research on Socio-Economic Policy (RESEP). Department of Economics, University Of Stellenbosch.

Glennerster, R. & Takavarasha, K. (2013). *Running randomised evaluations, A practical guide*. Princeton, NJ: Princeton University Press.

Glewwe, P., Kremer, M. & Moulin, S. (2007). Many Children Left Behind? Textbooks and Test Scores in Kenya.National Bureau of Economic Research Working Paper Series 13300. http://www.nber.org/papers/w13300.

Gove, A., & Wetterberg, A. (2011). *The Early Grade Reading Assessment: Applications and Interventions to Improve Early Grade Literacy.* RTI Press.

Grady, C. (2008). Clinical Trials, in *From Birth to Death and Bench to Clinic: The Hastings Center Bioethics Briefing Book for Journalists, Policymakers, and Campaigns*, ed. Mary Crowley (Garrison, NY: The Hastings Center), 21-24. http://www.thehastingscenter.org/Publications/BriefingBook/Detail.aspx?id=2164. Accessed 16 February 2016.

Groth-Marnat, G.(2009). Handbook of Psychological Assessment (Fifth ed.) Hoboken (NJ): Wiley.

Gustafsson, M. (2007). Using the hierarchical linear model to understand school production in South Africa. Stellenbosch Economic Working Papers, 01(07) http://resep.sun.ac.za/index.php/research-outputs/stellenbosch-working-papers/wp2007/ . Accessed 16 March 2016

Gustafsson, M. & Patel, F. (2008). Managing the teacher pay system. What the local and international data are telling us. Stellenbosch Economic Working Papers, 26(06).

Hanushek,E. & Woessmann,L. (2007). *The Role of Education Quality in Economic Growth*. (Policy Reseach Paper 4122), World Bank, Washington D.C.

Harrous-Paicheler, G. (undated).From data to evidence: Male circumcision as HIV prevention between controversies and scientific investigation. CNRS. CERMES3 UMR 8211.

Hausmann,R., Rodrik, D. & Velasco, A. 2006. Getting the Diagnosis Right. *Finance & Development* 43(1).

He, F., Linden, L & MacLeod, M. (2009). A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial.

Hellman, L. (2012). GPLMS Intersen Catch-up Programme: Analysis of Results. Memo.

Hopkins, W.G. (2000). Quantitative Research Design. Department of Physiology and School of Physical Education. http://www.sportsci.org/jour/0001/wghdesign.html. Retrieved 10 April 2011.

Human Sciences Research Council. (2013). Highlights from TIMSS 2011: South Africa. Pretoria: HSRC.

Imbens,G.W & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *Elsevier*. 142,2: 615–635. Retrieved 25 August 2015. www.elsevier.com/locate/jeconom.

International Initiative for Impact Evaluation (3ie). www.3ieimpact.org. Accessed 02 March 2016.

J-Pal. (2012). Evidence-Based Education. Policy Making and Reform in Africa Conference. Abdul Latif Jameel Poverty Action Lab. Accra, Ghana. May.

Khandker,S.R., Kodwa, G.B. & Samad, H.A. (2010) Handbook on Impact Evaluation Methods: Quantitative Methods and Practices. The World Bank. Washington DC.

Kremer, M., Brannen, C. & Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. Science, 340, 297. Retrieved 10 November 2015, www.sciencemag.org.

Lemons, C. J., Fuchs, D., Gilbert, J. K. & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242-252.

McCarney,R., Warner,J., Lliffe,S., van Haselen, R., Griffin, M,, Fisher, P., (2007). The Hawthorne effect: A randomised Controlled Trial. BMW Medical Research Methodology, 7, 30. iMed Central. The Open Access Publisher

McEwan, P. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, 85, 3: 353–394. http://rer.aera.net. Accessed 18 September 2015.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99: 210-221. www.elsevier.com/locate/devec. Accessed 24 July 2015.

Mckenzie, D. (2015) Notes from the AEAs: Present bias 20 years on + Should we give up on S.D.s for Effect Size? http://blogs.worldbank.org/impactevaluations/notes-aeas-present-bias-20-years-should-we-give-sds-effect-size. Accessed 5 September 2016.

Miguel, E & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the presence of treatment externalities. *Econometrics*, 72, 1: 159-217

Misra,S. (2012). Randomized double blind placebo control studies, the "Gold Standard" in intervention based studies. *Indian Journal of Sexually Transmitted Diseases and AIDS Journal*, 33(2): 131-134. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3505292/. Accessed 12 February 2016

Mouton, J. (2009). Assessing the impact of complex social interventions. *The Journal of Public Administration*, 44.

National Assessment of Educational Progress (NAEP). (2015). Item Mapping Procedures. http://nces.ed.gov/nationsreportcard/tdw/analysis/describing_itemmapping.asp . Accessed 5 September 2016.

National Education Evaluation and Development Unit. NEEDU National Report 2013: Teaching and Learning in rural primary school. Pretoria: Department of Basic Education

Piper, B. (2000) Integrated Education Program Impact Study of SMRS Using Early Grade Reading Assessment in Three Provinces in South Africa. United States Agency for International Development, and RTI International.

Piper, B. & Zuilkowski, S. (2015) Teacher coaching in Kenya: Examining instructional support in public and nonformal schools. *Teaching and Teacher Education*, 47: 173-183.

Pfost, M., Hattie, J., Dörfler, T. & Artelt, C. (2014). Individual Differences in Reading Development: A Review of 25 Years of Empirical Research on Matthew Effects in Reading. *Review of Educational Research*, 84, 2: 203-244. Retrieved 14 June 2015. http://rre.aera.net.

Pretorius, E. (2014). Supporting transition or playing catch-up in Grade 4? Implications for standards in education and training *Perspectives in Education.* University of Free State. http://www.perspectives-in-education.com 32(1)

Pritchett,L . & Beatty, A .(2012). The Negative Consequences of Overambitious Curricula in Developing Countries. CDG Working Paper 293 April 2012. Washington, D.C.: Center for Global Development. http://www.cgdev.org/content/publications/detail/1426129

Schochet, P.Z. (2005). Statistical Power for Random Assignment Evaluations of Education Programs. Mathematica Policy Research Inc. Institute of Education Sciences. Washington, DC.

Singh, A. (2015). How Standard is a Standad Deviation? A Cautionary Note on Using SDs to Compare across Impact Evaluations in Education. http://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations . Accessed 4 September 2016.

Spaull,N. (2012). Poverty & Privilege: Primary School Inequality in South Africa. Stellenbosch Economic Working Papers, 13(12). http://resep.sun.ac.za/index.php/research-outputs/stellenbosch-working-papers/wp2012/ Accessed 16 March 2016.

Spaull, N. (2013). South Africa's Education Crisis: The Quality of Education in South Africa 1995-2011. Johannesburg: Centre for Development and Enterprise.

Spaull, N. (2014). Education Quality in South Africa And Sub-Saharan Africa: An Economic Approach. Unpublished doctoral thesis, University of Stellenbosch, Stellenbosch.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, N., Schmidt, J., Jobse, H., Geelen, M., Pastorello, M. & Eyers, J. (2015). Interventions for improving learning outcomes and access to education in low- and middle-income countries: a systematic review. International Initiative for Impact Evaluation (3ie).

Svensson, J. & Pettersson-Lidbom, P. (2008). Impact Evaluation of Service Delivery Programs: Methods and Concepts for Impact Assessments in Basic Education, Health, Water and Sanitation. *African Economic Research Consortium*.

Taylor, S. & Watson, P. (2013). Developing and Evaluating the First Phase of the Grade 12 Mind the Gap Study Guide Series. Technical Report. Pretoria: Department of Basic Education.

Taylor, S. (2014). Unpublished Evaluation Tutorial document. Department of Basic Education.

Taylor, S. & Spaull, N. (2015). Measuring access to learning over a period of increased access to schooling: The case of Southern and Eastern Africa since 2000. *International Journal of Educational Development*, 41: 47–59. www.elsevier.com/locate/ijedudev.

Taylor, S. (2015). Measuring the impact of educational interventions. Retrieved 22 August 2015 from http://resep.sun.ac.za/index.php/events/conferences-2/#sthash.JpKuBBVE.dpuf

The Takshashila Blog. (2015).The Third Caveat of Randomised Control Trials: Intra-hypothesis validity. www.logos.nationalinterest.in/2015/03

University of Pretoria. (2012). PIRLS 2011: South African Children's Reading Literacy Achievement Report. Centre for Evaluation and Assessment, University of Pretoria. Retrieved 12 December 2012 from www.up.ac.za.

Van der Berg, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. Centre for European Governance and Economic Development Research. Discussion paper 69. January.

Van der Berg, S. (2014). Education policy and quantitative research in SA. Unpublished presentation.

Venkatakrishnan,H. & Spaull,N. (2014). What do we know about primary teachers' mathematical content knowledge in South Africa? An analysis of SACMEQ 2007. Stellenbosch Economic Working Paper.

World Bank. (2010.) Impact Evaluations in Practice. Retrieved 10 October 2014 from http://www.worldbank.org/ieinpractice

Zani, B., Pienaar, E., Oliver, J. & Siegfried, N. (2011). Randomized Controlled Trials of HIV/AIDS Prevention and Treatment in Africa: Results from the Cochrane HIV/AIDS Specialized Register. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028759.

Zuze, L. & Reddy, V. (2014). School resources and the gender reading literacy gap in South African schools. International Journal of Educational Development 36: 100–107. www.elsevier.com/locate/ijedudev

APPENDIX A: Declaration - Article

eVALUAtion Matters

A Quarterly Knowledge Publication on Development Evaluation

Second Quarter 2015

IMPACT EVALUATION:
Insights from Practitioners


Emmanuel Jimenez


Richard Manning


Annette N. Brown


Lycia Lima


André Portela Souza


Dalila Figueiredo


Elsa de Morais Sarmento


Christian Kingombe


Laura Camfield


Maren Duvendack


Richard Palmer-Jones


Jean-Louis Arcand


Jyotsna Puri


Cameron Breslin


Alexandra Orsola-Vidal


Temina Madon


Jacob Oduor

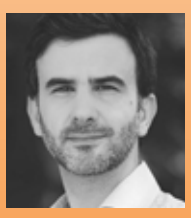

Laurence Lannes


Fabrizio Felloni


Simona Somma


Nompumelelo Mohohlwane


Stephen Taylor


Ricardo Paes Mamede


Daniela Cruz


Teresa Farinha Fernandes


Jeremy D. Foltz


Kweku Opoku-Agyemang


Peter Huisman


Rik Linssen


Anne Oudes

AFRICAN DEVELOPMENT BANK GROUP
GROUPE DE LA BANQUE AFRICAINE DE DÉVELOPPEMENT

IDEV
Independent Development Evaluation
African Development Bank

idev.afdb.org

Contents

… poor people's lives can be improved if the development community learned more systematically from its efforts – in particular, if more rigorous impact evaluations of what works in development were done, if their results were made widely available and understood and if policymakers and program managers used that evidence to improve policy and practice.

Emmanuel Jimmenez, p. 8



4 **From the Evaluator General's Ipad**

Impact evaluation is a complement to existing methodologies that are more focused on organizational effectiveness. It is an additional tool that allows evaluation to showcase itself as truly independent in assessing development effectiveness.

8 **Interview: Emmanuel Jimenez**

Impact evaluations of development programs are on the rise, compared to a decade ago. What are the emerging trends and what needs to be done to institutionalize IE in Africa?

12 **Interview: Richard Manning**

… good evidence from effective impact evaluations and systematic reviews has to compete in a political space where there is pressure or decisions to be taken that are expedient in terms of publicity and electoral cycle, whether or not they are based on evidence.

14 **Interview: Annette N. Brown**

Institutions that are seeking to innovate and leapfrog ahead in IE should start small, learn by doing and stay involved.

18 **Impact Evaluation of Development Programs: Challenges and Rationale**
Lycia Lima, André Portela Souza, Dalila Figueiredo

Why policymakers should conduct an impact evaluation of their programs and the main challenges they may face in doing so.

24 **Improving Impact Evaluation in Development Programs/Projects**
Elsa de Morais Sarmento, Christian Kingombe

How to improve impact evaluation usage in development programs or projects.

31 **Bias in Evaluation**
Laura Camfield, Maren Duvendack and Richard Palmer-Jones

Qualitative and quantitative evaluations are vulnerable to political, social, and economic pressures; the personalities of evaluation researchers frame what they "see" and what it is possible for them to see, and what they strive to find statistical support for. An overview of empirical, researcher, methodological, and contextual bias.

40 **Randomization in the Tropics: The Earnest Search for a Mirage?**
Jean-Louis Arcand

For those who have spent many years working in the field attempting to convince policymakers in developing countries to actually evaluate the effectiveness of their programs, RCTs have, to a large extent not lived up to their promise.

46 **Embracing Complexity: Impact Evaluation-Ready Systems – Some Experiences**
Jyotsna Puri

Challenges in implementing impact evaluations include lack of awareness about the potential power and use of impact evaluations; inadequate data to inform impact evaluations;and little engagement by policy makers. Despite these, impact evaluations have constantly evolved to teach us important lessons about implementing them and in understanding policy. Some examples are discussed.

eVALUatiOn Matters
is a quarterly publication from Independent Development Evaluation at the African Development Bank Group. It provides different perspectives and insights on evaluation and development issues.

Editor-in-Chief: Felicia Avwontom (f.avwontom@afdb.org)
Produced under the guidance of the AfDB Evaluator General, Rakesh Nangia.

This issue was co-produced with Elsa De Morais Sarmento as guest editor.

Acknowledgments: IDEV is grateful to all contributors, reviewers, editors, and proofreaders who worked on this issue.

© 2015—African Development Bank (AfDB)

African Development Bank Group
Immeuble du Centre de commerce international d'Abidjan (CCIA)
Avenue Jean-Paul II
01 BP 1387, Abidjan 01
Cote dIvoire
Phone: +225 20 26 44 44
Fax: +225 20 21 31 00
Internet: www.afdb.org

Design: Phoenixdesignaid and Felicia Avwontom
Layout: Phoenixdesignaid

56 **An Innovative Three-pronged Approach to Institutionalise Impact Evaluation Within Multilaterals**
Cameron Breslin, Alexandra Orsola-Vidal, Temina Madon

Despite the need and demand for rigorous evidence, impact evaluation has yet to become fully institutionalized within multilateral development agencies. This article describes a program designed to build capacity and promote the mainstreaming of impact evaluation within multilaterals.

64 **Impact Evaluation in Development Institutions: A Look at the African Development Bank**
Jacob Oduor, Elsa de Morais Sarmento, Laurence Lannes

The Bank only recently started conducting rigorous research on what actually works and what can be attributable to its own interventions. Achievements and remaining challenges.

70 **Impact Evaluation Matters: Enhanced Learning Through Involving Stakeholders in Oxfam's Impact Studies**
Peter Huisman, Rik Linssen, Anne Oudes

An overview of the World Citizens Panel developed by Oxfam as a new approach to impact measurement. Oxfam has implemented the approach in six countries in Asia and Africa and preparations in eight other countries are in progress.

77 **Impact Evaluations in Rural Development: Opportunities and Challenges. The Emerging Experience of IFAD's Independent Office of Evaluation**
Fabrizio Felloni, Simona Somma

Impact Evaluation at The Independent Office of Evaluation: past, present and future.

90 **Steps in the Impact Evaluation of AfDB-Financed (Rural) Road Projects**
Christian Kingombe

On the road to successful implementation of a (rural) road impact evaluation: Six interlinked steps that could constitute the basis of an impact evaluation user guide.

99 **Using Impact Evaluation for Education Policy Innovations: the Case of Early Grade Literacy in South Africa**
Nompumelelo Mohohlwane and Stephen Taylor

In the absence of solid evidence of effective policies and programs, the use of prospective impact evaluations is recommended.

108 **Evaluating the Impact of Vocational Education in Portugal: a Successful Case of Institutional Collaboration**
Ricardo Paes Mamede, Daniela Cruz and Teresa Farinha Fernandes

A counterfactual evaluation of the impact of vocational courses on students' school and labor market performance.

114 **Evaluations Can Help Policymakers Understand Corruption**
Jeremy D. Foltz, Kweku Opoku-Agyemang

How impact evaluations can help us understand petty corruption as a phenomenon and help remove some of the stigma associated with discussing the problem in a non-superficial way.

> *Progress toward the institutionalisation of evaluation can be demonstrated through: Country-led initiatives; Strong "buy-in" from key stakeholders; and Evaluation capacity development.*
>
> *Breslin et al., p. 62*





The **mission of Independent Development Evaluation at the AfDB** is to enhance the development effectiveness of the institution in its regional member countries through independent and instrumental evaluations and partnerships for sharing knowledge

Evaluator General: Rakesh Nangia, r.nangia@afdb.org

Managers:
Samer Hachem, S.hachem@afdb.org
Karen Rot-Munstermann, k.rot@afdb.org
Rafika Amira, r.amira@afdb.org

Questions?
Telephone: +225 2026 2041

Web: http://idev.afdb.org
Write to us: EvaluationMatters@afdb.org
idevhelpdesk@afdb.org

Copyright: © 2015—African Development Bank (AfDB)





Nompumelelo Mohohlwane and Stephen Taylor

Using Impact Evaluation for Education Policy Innovations: *the Case of Early Grade Literacy in South Africa*

20 March 2015

Background: Development through quality education

South African and international development planning is increasingly centered around education. Agreements such as the Millennium Development Goals (MDGs) and the Education For All (EFA) drive are testament to the international recognition of the pivotal role of education. In South Africa, the National Development Plan, which is arguably the country's most prominent planning document, affords a central role to education, while the Presidency officially regards improved quality of basic education as the country's number 1 priority (NPC, 2012).

Despite substantial progress in expanding access to schooling in developing countries over the past few decades, there is now a growing recognition that in many countries the learning outcomes achieved by those attending school are often dismally poor. Spaull and Taylor (2015), for example, demonstrate that despite improved access to schooling in many Southern and East African countries there are large proportions of children who reach grade 6 without having acquired basic literacy and numeracy skills. This is important since there is clear evidence that the quality of skills achieved (over and above the quantity of schooling attained) has a significant impact on economic growth and on the labour market prospects of individuals (Hanushek & Woessmann, 2007). While there are numerous social, psychological and other benefits of education, the quality of learning outcomes should be seriously considered when analyzing education system performance.

The education quality challenge in South Africa

Local and international assessments of learner performance consistently indicate poor performance by South African learners in Mathematics, Science, and Languages (Spaull, 2014). The low performance across the board is concerning but for the purposes of this paper the focus will be on literacy and reading.

The Pre-PIRLS 2011 results indicated that 29 per cent of South African Grade 4 learners did not have the rudimentary reading skills required at a Grade 2 level. The situation was most severe for those learning in an African language. For example, 57 per cent of learners that took tests in Sepedi or Tshivenda did not reach this level (University of Pretoria, 2012). The 2006 PIRLS study, which tested grade 5 South African learners on a somewhat more advanced reading test, showed that approximately 80 per cent of children had not learned to read for meaning by grade 5, but at best could only extract basic factual details from a text. If children have not learned to read fluently by this time, it stands to common sense that they will not be able to cope with the requirements of the curriculum at higher grades. Weak reading foundations are therefore at the heart of the education quality challenge in South Africa.

South Africa's underperformance, even relative to many poorer countries, prompts an enquiry into the causes behind the limited ability of the education system to convert inputs into outcomes. South Africa's per pupil expenditure in schools in purchasing power parity (PPP) terms exceeded that of all the other thirteen countries in SACMEQ with the exception of Seychelles. An important part of the contextual background is the inequality resulting from the political history of South Africa. The changes in the education system following the end of Apartheid and the establishment of a new democratic state in 1994 have had

limited success in changing the nature of schooling. Several authors, such as Fleisch (2008) and Spaull (2014), argue that South African schooling has still effectively got two systems in one. The first and largest part of the system comprises the historically disadvantaged schools and is characterized by inefficiency including poor school management, continuous underperformance, high and indiscriminate grade repetition and dropout. The second system refers to historically white and Indian schools where learner performance is at a higher standard, parents make substantial fee contributions, organizational and instructional processes are more efficient and schools are well endowed with infrastructure.

The South African government is well aware of these challenges and continues to allocate the largest share of government expenditure to education. Since the early 1990s education spending has become increasingly well targeted to poor schools (Gustafsson and Patel, 2008). Specific initiatives and policies implemented by the South African government to address equity challenges in education include the introduction of no-fee schooling which is implemented in about 77 per cent of public schools, and the provision of daily meals through the National School Nutrition Program to approximately 70 per cent of schools focusing on the poorest schools (Department of Basic Education, 2014).

Despite these considerable efforts, however, learning outcomes remain low in South Africa and little is known about the effectiveness of particular policies and programs designed to improve learning. Where evidence is available it is often self-reported, focused on inputs, anecdotal or part of a larger initiative where the effect of specific efforts is difficult to isolate. This motivates for an agenda of impact evaluation to inform policy-making going forward, as will be argued below.

The importance of early literacy learning in South Africa

Literature on the evidence of early learning emphasizes the importance of mastering certain learning foundations for the sake all further learning. The literature refers to 'self-productivity', explaining that skills acquired during one period generally persist into the next period and may make the acquisition of other skills in another dimension easier (Girdwood, 2013).

In addition to the argument for the cognitive benefits of the development of good educational foundations and their lasting effects, James Heckman (2007), amongst others, contends that intervening earlier rather than later is more cost-effective. The costs of providing curriculum support for areas of learner deficits identified early, such as in the Foundation Phase are expectantly lower than mediating learning later in schooling where the gap between curriculum expectations and learner knowledge may be excessively large in a multitude of subjects, as Pritchett and Beatty (2015) have shown. The costs accrued at later stages include high rates of grade repetition and dropping out of the education system.

One critical learning foundation that needs to be acquired during the early grades of primary schooling is reading. A large theoretical literature points to the benefits of learning to read in the home (or first) language. One of the expected benefits is that second language acquisition should be easier once a firm grasp of the nature of reading and literacy has been attained in one language. A paper by Taylor and Coetzee (2013) provides some empirical evidence from South Africa that home language instruction in grades 1 to 3 caused improved English literacy in grades 4 to 6 compared with children who were taught in English as the language of instruction. This finding substantiates the argument that all learning builds on prior learning; as such mastery of a second language is enabled by the mastery of the first language. This points to the strategic value of finding ways to improve home language reading acquisition in the Foundation Phase. Yet, the reality is that the majority of children will experience a transition to English as the language of instruction in the fourth grade. Finding ways to strengthen English vocabulary and manage this transition most effectively will therefore also be important.

Education Policy Development and the Evaluation Process

A detailed review of South African education policy development is beyond the scope of this paper. What is clear, however, is that despite many policy changes and new programs, little is known about the ultimate impact of these initiatives on learning outcomes. The lack of a focus on impact evaluation is not unique to South Africa, as the following quote illustrates:

"Development programs and policies are typically designed to change outcomes, for example, to raise incomes, to improve learning, or to reduce illness. Whether or not these changes are actually achieved is a crucial public policy question but one that is not often examined. More commonly, program managers and policy makers focus on controlling and measuring the inputs and immediate outputs of a program—how much money is spent, how many textbooks are distributed—rather than on assessing whether programs have achieved their intended goals of improving well-being" (World Bank, 2010).

In cases where extensive research is done – at least in South Africa – it is typically focused on diagnosing areas requiring attention rather than evaluating possible solutions. Where interventions are evaluated it is often through conducting case studies or piloting in a small number of schools. The shortcoming of this approach is that the implementation model often used in case studies or small-scale pilots is often resource intensive and may be difficult to replicate at a larger scale.

A focus on evaluation is now emerging within the South African government through the introduction of the National Evaluation Policy Framework in 2011. This policy framework includes a National Evaluation Plan (NEP) which commissions independent evaluations of priority government programs in a partnership between the custodian department and the Department of Planning, Monitoring and Evaluation (DPME, 2014). Several Department of Basic Education (DBE) programs have been evaluated through the NEP, namely the Grade R program, the Funza Lushaka Bursary Program and the National School Nutrition Program.

The evaluations referred to above are all retrospective evaluations, assessing how well programs were implemented or if the intended program goals were attained. Prospective impact evaluations, where programs are evaluated prior to being taken to scale, remain extremely rare. One exception to this is the impact evaluation of a new set of study guides introduced by the DBE in 2012 (Department of Basic Education, 2013).

Using Randomised Control Trials in Education

The major challenge in impact evaluation is the need to identify a counterfactual – what would have happened to program recipients in the absence of the intervention? Since one can never actually observe a counterfactual to reality, one needs to use a "control group" or "comparison group" to provide a valid estimate of the counterfactual. Simply comparing recipients with non-recipients or pre- and post-outcomes amongst recipients is usually not likely to provide a valid estimate of the counterfactual since recipients are usually systematically different to non-recipients and outcomes would change over time in any event.

While various quantitative impact evaluation methods are available, the cleanest method for identifying an internally valid estimate of the counterfactual is obtained through conducting a Randomised Controlled Trial (RCT). Through using a lottery to allocate participants to intervention and control groups, an RCT constructs a credible "counterfactual" scenario – what would have happened to those who received an intervention had they not received that intervention.

Prospective impact evaluations also have the advantage for research of uncovering knowledge of the binding constraints in the school system. In complex environments, such as education, there are multiple factors influencing outcomes and it is not always clear which factors to address first. For example,

high quality teaching requires both competent and motivated teachers, though it is not clear which of these is the more binding constraint in South Africa. Teacher knowledge in South Africa is weak: Carnoy et al (2011) found that grade 6 teachers recorded an average score of around 40 per cent on a test designed to assess their mathematics knowledge for that grade. Furthermore, studies show low teacher motivation in poor schools, manifested in high absentee rates and low teaching activity (Reddy et al, 2010). Yet, it is unclear whether to address teacher capacity or teacher motivation first. The lack of rigorous evaluations to establish which of these challenges to address first is a shortcoming of conventional policy and program development.

Practical considerations when implementing an RCT

Statistical expertise is required in the design of an RCT. This involves calculating the required sample sizes in each intervention and control group and conducting the random assignment. For practical reasons, when conducting an education RCT it is often necessary to assign schools as a whole to intervention or control groups, as opposed to assigning individuals to the different experimental groups. This leads to rather large required samples, which has cost implications. The need to raise funds takes time and requires significant stakeholder engagement and government support to convince donors to be involved.

There are two main components to an education RCT – there is the *implementation* of the new interventions and there is the *evaluation* of their impact. The evaluation side of the RCT involves the collection of outcomes data as well as contextual data for the sake of measuring changes in intermediate outcomes and identifying factors that mediate the impact of the intervention. Both the implementation and the evaluation components require financing and should be conducted by separate organizations. In some cases, an NGO or a government department may fund and implement the interventions, thus reducing the need for additional fund raising. If reliable outcomes data already exist, through for instance a nationally standardized examination, then one might be able to significantly reduce costs associated with the evaluation side of the project.

Both academic researchers and implementing agencies face various perverse incentives when considering or conducting evaluations. A publication bias exists in academia where it is more likely to see studies with positive results published than studies showing no impact (Duflo, Glennerster and Kremer, 2006). RCTs are less prone to this bias, since the large investment of time and resources together with the high reliability of the results mean that even evaluations showing zero impact are likely to be published. Government departments and NGOs may resist evaluations due to the risk of negative findings. Therefore, prospective evaluations of alternative programs or variations of programs under consideration may be more amenable to policy makers and program managers who will then not feel that their entire work for several years is being judged.

For these reasons, the success of any RCT is dependent on extensive stakeholder consultation and support. This ensures that funds and other resources including personnel are availed; the integrity of the research design is upheld; the implementation of the interventions is conducted properly, and the findings are considered for program or policy scale-up or redesign.

Experiences from two new reading evaluations in SA

Through a developing partnership between education researchers, government and donors, two studies are being undertaken to evaluate possible ways to improve reading acquisition in South African schools.

Evaluation of a remedial reading program in Grade 4

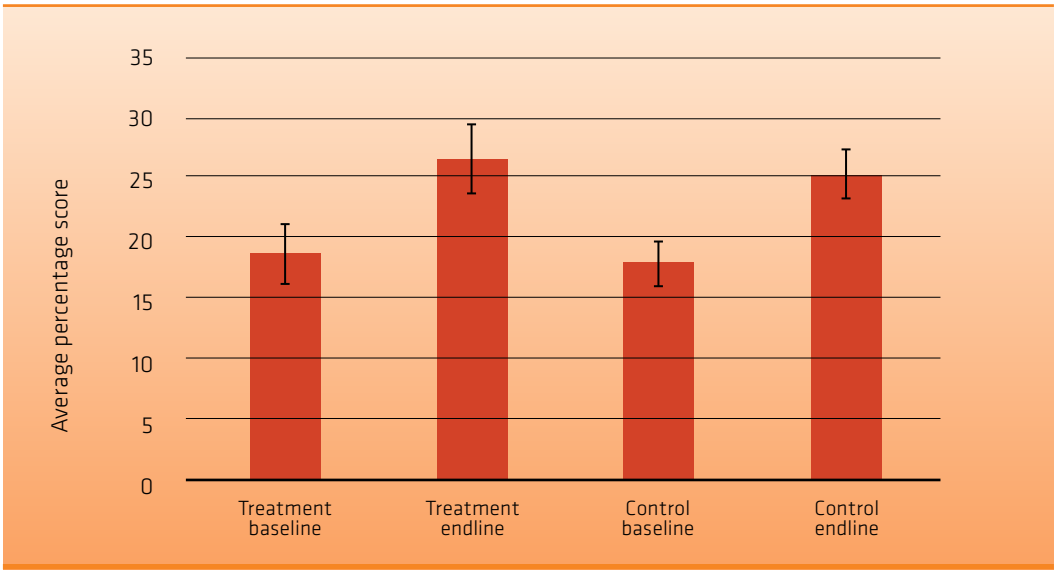
The Gauteng Primary Language and Mathematics Strategy (GPLMS) implemented between 2011 and 2014, included various new interventions focused on the early grades. As part of this, a Reading Catch-Up Program (RCUP) was developed to strengthen the

English skills of children in grades 4 and 5 whose first language was not English but who are required to learn using English as the language of instruction in those grades. A preliminary evaluation of the program indicated large gains in the language skills of program recipients over time (Hellman, 2012). However, there was no control group. Therefore, some strong assumptions had to be made about how much learning would have taken place over the period had there been no intervention.

These initially promising, though inconclusive, results prompted an RCT of the RCUP to be conducted in the district of Pinetown in the KwaZulu-Natal province in 2014 (Fleisch, Taylor, Schöer, and Mabogoane, 2015). The intervention lasted for 11 weeks and consisted of on-site teacher support by reading coaches and the provision of scripted lesson plans and additional graded reading books. The RCUP targeted Grade 4 learners in schools that transition to English as the language of instruction after using the home language in the Foundation Phase. The hypothesis underlying the program was that the learning gaps in learner mastery of English at the end of the second term in Grade 4 may be caught-up through the provision of a well-designed relatively short intervention. The program was implemented in 40 intervention schools with a control group of 60 schools. Assignment to intervention and control group was done through a computerized lottery. Different organizations were contracted to conduct the intervention and the data collection for evaluation, and the evaluation agent was kept blind to which schools were in the intervention group versus the control group.

The most notable finding of the study was that although learners in intervention ("treatment") schools improved their test scores between the baseline and the endline assessment, the learners in control schools improved by a similar margin, as depicted in Figure 1. This illustrates the importance of obtaining an estimate of the counterfactual: in the absence of a randomly selected control group a false positive result would have been obtained. The main finding, then, is that the RCUP intervention had no statistically significant impact on the overall reading achievement of learners. However, treatment schools improved more than control schools in the spelling and grammar subcomponents of the test. The program impact was larger for learners who initially had a basic minimum of English skills and for those whose teachers participated actively in the program.

Figure 1 Average Pre- and Post-Scores for Intervention and Control Schools



Note: 95% confidence intervals are indicated
Source: Fleisch, Taylor, Schöer, and Mabogoane, 2015

The findings from the RCUP RCT yield several important policy lessons and raise several questions for further research. Firstly, it is evident that ostensibly well-designed programs may not have as large an effect as one might expect in the absence of a rigorous evaluation. Apart from the design quality of a program and the integrity of implementation, there may be contextual factors pertaining to learners, schools and communities which either preclude or are conducive to the effectiveness of an intervention.

Secondly, the findings indicate that the RCUP cannot yet be implemented on a wider scale with any confidence that it will have a significant impact on learning outcomes. This does not necessarily mean that reading remedial programs in general, or even the RCUP specifically, should not be further explored. It does, however, mean that a revised version should first be experimented with and shown to work before government should consider a larger scale implementation.

Thirdly, the independently administered tests indicated that the learning deficits existing by the end of the Foundation Phase are apparently much larger than expected. The finding that initially better-performing learners gained more from the intervention may imply that the program would have been more appropriate at the grade 5 level, even though it was covering topics that should have been covered in the Foundation Phase. Another possibility is that an 11-week intervention is simply too short a time to deeply influence classroom practice and learning.

The Early Grade Reading Study (EGRS)

The second RCT which is being conducted by the Department of Basic Education (DBE) is the Early Grade Reading Study (EGRS). This RCT has recently (February, 2015) commenced in 230 schools in the North West province and has been crafted to evaluate three competing interventions all aimed at improving home language (Setswana) reading acquisition in grades 1 and 2.

The sample of schools includes non-fee paying schools in the Dr Kenneth Kaunda and Ngaka Modiri Molema districts of the North West province. All schools selected use Setswana as the language of instruction in the Foundation Phase. All three interventions will occur over a two-year period working with the cohort of children entering grade 1 in 2015. The RCT will evaluate the causal impacts of three interventions: (i) a teacher training course focused specifically on the teaching of Setswana reading and literacy, accompanied by scripted lesson plans and graded reading materials; (ii) an on-site support program to teachers from reading coaches, accompanied by scripted lesson plans and graded reading materials; (iii) and a package designed to improve parent involvement in – and monitoring of – learning to read. Each intervention will be implemented in 50 schools within the sample. A further 80 schools have been selected as the comparison group. As was the case in the RCUP study, separate organizations have been contracted to undertake the implementation of interventions and the data collection for evaluation, with the evaluation agent blind to which group schools fit into.

This study is expected to shed light on several research and policy questions. Firstly, it will show which of three alternative interventions is most cost-effective. Although each intervention has a different unit cost, the improvement in test scores per Rand spent for each intervention will be calculated. The evaluation will also investigate whether the impacts of interventions are different for various sub-groups of learners or schools. This will inform the most appropriate targeting of interventions if scaled up. The study is also designed to look at long-term effects and spillover benefits of faster reading acquisition. Do the impacts of the interventions persist, dissipate or compound over time? If one succeeds in improving the acquisition of home language reading in the early grades, are there spillover benefits into other learning areas such as Numeracy and First Additional Language? This will be measured using results of the Annual National Assessments (ANA) in subsequent years.

Conclusion

This paper has demonstrated that improving the acquisition of reading in the early grades is central to the education quality challenge in developing countries, and especially in South Africa. In the absence of solid evidence of effective policies and programs to address this challenge, the use of prospective impact evaluations is recommended. The paper has pointed out the necessity for innovation and rigor to establish such evidence and understanding the binding constraints in the complex South African education system.

These points are substantiated through a description of two recent RCTs focusing on early grade reading in South Africa. The RCT of the Reading Catch-up Program, implemented in Pinetown, Kwa-Zulu Natal has provided important lessons. The findings have highlighted the need for a valid counter-factual in measuring impact, which is a strength of the RCT methodology. The second RCT discussed, the Early Grade Reading Study (EGRS), which is being implemented in 230 schools in the North West province provides an exciting opportunity for further learning. It is anticipated that the findings will address some of the questions emerging from the RCUP study as well as provide substantive information on the binding constraints in the teaching of language in South African schools.

References

Boiseiere, M. (2004). "*Determinants of Primary Education Outcomes in Developing Countries*". Washington: The World Bank.

Carnoy, M. Chisholm, L. Addy,N. Arends,F. Baloyi, H. Irving, M. Raab, E. et al. (2011). *The process of learning in South Africa: the quality of mathematics teaching in North West Province.* Technical Report. Stanford University School of Education and HSRC.

Department of Basic Education. (2014). *Annual Report 2013/2014.* Pretoria. Retrieved 15 May 2013 from www.education.gov.za.

Department of Basic Education (2013). Technical Report. *Developing and Evaluating the First Phase of the Grade 12 Mind the Gap Study Guide Series.* Pretoria.

Department of Basic Education. (2012). *Summary Report: Evidence Based Policy Review Workshop.* Pretoria.

Department of Basic Education. (2014). Terms of Reference. *Implementation of Reading Interventions for the research project: "Early Grade Reading Study.* Pretoria.

Department of Planning, Monitoring and Evaluation. (2014). *Concept for the National Evaluation Plan* 2015/16. Pretoria.

Duflo, E., Glennerster, R., & Kremer, M. (2007). *Using Randomization in Development Economics Research: A Toolkit.* Center for Economic Policy Research.

Fleisch, B. (2008). Primary Education in Crisis: *Why South African schoolchildren underachieve in reading and mathematics.* Cape Town: Juta.

Fleisch, B.Taylor, S. Schöer, V. & Mabogoane, T. (2015). *Assessing the impact of the RCUP: The findings of the impact evaluation of the Reading Catch-Up Program.* Paper Presented at the Wits School of Education Research Seminar.

Girdwood, E. (2013). *Exploring the Evidence on Early Learning: A Literature Review.* Impact Evaluation of the Introduction of Grade R on Learning Outcomes. Research on Socio-Economic Policy (RESEP). Department of Economics, University Of Stellenbosch.

Gustafsson, M. & Patel, F. (2008). *Managing the teacher pay system.* What the local and international data are telling us. Stellenbosch Economic Working Papers. 26(06).

Hanushek,E,A & Woessmann,L.(2007). *Education quality and economic Growth.* The World Bank: Washington.

Heckman, J. (2007). *The economics, technology, and neuroscience of human capability formation.* NBER Working paper 13195.

Hellman, L. (2012). GPLMS Intersen Catch-up Program: Analysis of Results. Memo.

National Planning Commission. (2012). *National Development Plan: Vision for 2030.* Retrieved January 18, 2013, from http://www.npconline.co.za/.

Pritchett,L. & Beatty,A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development.* 40 (1), 276–288.

Reddy, V., Prinsloo, C., Netshitingani, T., Moletsane, R., Juna, A., & Janse van Rensburg, D. (2010). *An investigation into educator leave in the South African ordinary public schooling system.*

Spaull, N. (2014) *Education Quality In South Africa And Sub-saharan Africa: An Economic Approach* (Unpublished doctoral thesis, University of Stellenbosch, Stellenbosch).

Spaull, N., & Taylor, S. (2015). Access to What? Creating a Composite Measure of Educational Quantity and Educational Quality for 11 African Countries. *Comparative Education Review,* 59 (1), 133-165.

Taylor, S. & Coetzee, M. (2013). *Estimating the impact of language of instruction in South African primary schools: A fixed effects approach.* Stellenbosch Economic Working Papers: 21/13.

University of Pretoria. (2012). *PIRLS 2011: Media Release.* 11 December 2012. Retrieved 12 December 2012. from www.up.ac.za.

World Bank. (2010) *Impact Evaluations in Practice.* Retrieved 10 October 2014 from http://www.worldbank.org/ieinpractice.

PROFILE OF THE AUTHORS

Nompumelelo Mohohlwane is an Assistant Director in the Research Coordination, Monitoring and Evaluation Directorate at the South African Department of Basic Education. The unit is responsible for system monitoring, supporting performance information management, and conducting research and evaluation of education interventions. She is currently completing her Masters in Education at the University of Witwatersrand.
Affiliations: Department of Basic Education

Stephen Taylor is an Advisor in the South African Department of Basic Education. His work includes impact evaluation of education interventions, measuring educational performance and equity in educational outcomes. In 2010 he completed a PhD in economics at the University of Stellenbosch, analysing educational outcomes of poor South African children.
Affiliations: Department of Basic Education Research Associate of the Department of Economics, University of Stellenbosch

APPENDIX B: CHAPTER 4 RCUP Letter of Permission

Wits School of Education

Ggg

Permission to use the Reading Catch-Up Programme Study dataset

As the principal investigator of the Reading Catch-up Programme Study, this letter serves to confirm that Ms Nompumelelo Mohohlwane, a registered student for a Masters in Education with the University of Witwatersrand, is granted permission to use the final and confirmed dataset from the Reading Catch-Up Programme Study.

Please note that a full acknowledgement of the source of the data needs to be included in the Research Report. Full acknowledgement of the original study funder, the Zenex Foundation, must also be included. The acknowledgement must contain a disclaimer indicating the results, findings and conclusion contained in the Research Report are those of the author alone.



Prof Brahm Fleisch
Head of Division, Educational Leadership, Policy and Skills
Wits School of Education

4 February 2015

APPENDIX C: CHAPTER 4 RCUP Main Report

Assessing the impact of the RCUP: A report of the findings of the impact evaluation of the Reading Catch-Up Programme

Brahm Fleisch, Stephen Taylor, Volker Schöer, and Thabo Mabogoane

January 2015

Funding for this study has been provided by the Zenex Foundation.

Contents

Table and figures 3
Acknowledgements 4
1. Introduction 5
2. Research design and methods 7
2.1 Intermediate Phase Catch-Up Programme 7
2.2 Experimental design 10
2.3 Site description 11
2.4 Sampling frame and rationale 11
2.5 Pretest learner results 13
2.6 Implementation 15
3. Results 17
4. Discussion 29
5. Conclusion 32
References 33
Annexure 35
Endnotes 42

Table and figures

Table 1: Baseline performance by age of learners 14
Table 2: Baseline performance by gender 14
Table 3: Level of implementation and teacher competency 15
Table 4: Curriculum coverage 16
Table 5: Attrition between pretest and posttest, RCUP 2014 17
Table 6: Main regression results 19
Table 7: Impact by baseline performance of learners 20
Table 8: Teacher attendance at afternoon workshops 21
Table 9: IV estimates of the treatment effect on the treated 22
Table 10: The impact of the coaches 23
Table 11: Treatment effect on Grade 4 English First Additional Language (ANA) 25
Table 12: The effect of treatment on Grade 4 Mathematics (ANA) 26
Table 13: Treatment effect on language across untreated grades (ANA) 27
Table 14: The impact of the coaches (ANA) 27
Table 15: The impact of treatment on ANA language scores (for individually matched sample) 28

Figure 1: Catch-up programme distribution across the four levels, 2012 10
Figure 2: Kernel density of pretest scores, percentage 14
Figure 3: Distribution of baseline performance by gender 15
Figure 4: Posttest score distributions for treatment and control schools 18
Figure 5: Mean scores for treatment and control groups (pre- and posttest) 18
Figure 6: Kernel density curves of test scores for Grade 4 English First Additional Language (ANA, 2014) 25

Acknowledgements

We would like to acknowledge the financial support of the Zenex Foundation, without which this study would not have been possible. That said, the authors take full and sole responsibility for the views and content of this report. There are a number of other organisations and individuals that need to be acknowledged and thanked for their ongoing support and assistance. We would like to thank the Wits University Human Research Ethics Committee (Non-Medical) for its scrutiny and flexibility. To Mary Metcalfe, a big thank you for facilitating our discussions with the KwaZulu-Natal Department of Education. We need to thank the top management of that organisation for having the foresight to allow the study to take place. We owe a real debt of gratitude to the Pinetown district staff, and, above all, to the teachers, both in the control and the treatment schools, for their ongoing commitment. Finally, to our partners at Class Act and the JET Education Services, thank you for your sterling work.

1. Introduction

Over the past decade there has been a growing recognition that a substantial proportion of schoolchildren in South Africa are one or more years below the acceptable achievement levels, particularly in key subjects, such as English First Additional Language and Mathematics (Taylor, 2014; NEEDU, 2014; Spaull, forthcoming). Spaull (forthcoming) makes the compelling case that schoolchildren that are academically behind the acceptable levels of performance in the Foundation Phase are likely to fall further and further behind their counterparts as they progress up the school system. This is clearly not a conventional 'remedial' problem, that is, a small number of individuals in a class that have specific learning barriers or challenges, but rather the learning deficits are systemic, often affecting almost all learners in the majority of disadvantaged schools.

How can education departments address these systemic learning backlogs? There are a growing number of specialised programmes, particularly at the Grade 12 level, that focus on providing additional specialist instruction. Although the systemic achievement gap often begins at the Foundation Phase, fewer programmes have been developed specifically to address the systemic problem early in learners' school careers. One exception is the Intermediate Phase Catch-Up Programme, which was developed as a component of the Gauteng Primary Literacy and Mathematics Strategy in 2012. The eleven-week programme, which focused on reteaching Foundation Phase English First Additional Language skills and content to learners in underachieving primary schools, was designed to replace the curriculum for a single term, to ensure that learners in these schools had an opportunity to master the basics of English-language literacy. Hellman's (2012) interval evaluation showed that the Intermediate Phase Catch-Up Programme was effective at scale in helping the majority of learners in Grades 4 to 6 to gain basic literacy proficiency. But, while the results were clearly encouraging, the design of the internal evaluation was not rigorous. The impact evaluation was administered by the service provider that had designed the intervention, the pre- and posttest instruments were administered by the teachers themselves, and the study did not have a counterfactual component.

Given the importance in the education space of systemic catch-up programmes, and the need for robust evidence of their effectiveness, a research team designed a robust impact evaluation of the Gauteng Intermediate Phase Catch-Up Programme. Although evidence generated from small-scale pre- and posttest studies (such as Pretorius, 2014) has the potential to contribute to the knowledge base, there are clear advantages of randomised experiments. Randomised experiments allow researchers to establish with greater certainty the efficacy of education initiatives and/or specific programme interventions.

The impact evaluation of what came to be called the Reading Catch-Up Programme (RCUP)[1] had a number of design features to ensure robustness. The research team, which designed the study, and which analyses and reports on the findings, separated the study into a learner data-collection component and an implementation component. Class Act, the agency originally involved in the development of the intervention, was tasked with implementing the intervention in treatment schools.[2] JET Education Services was responsible for collecting learner information from pre- and posttests in both treatment and control schools. The intervention took place from April to June 2014 in Pinetown, KwaZulu-Natal.

The final report is structured into four sections. Following this brief introduction, the report provides a detailed description of the study method, focusing on a description of the intervention, the randomised control trial (RCT) methodology, the rationale for the selection of the study site, and the data-collection processes. The third section presents the major findings, including both information from a qualitative case study undertaken during the intervention and the results of the pre- and posttesting. While the focus here is on the main findings of the impact evaluation, this section also provides insights about other related findings. The discussion section of the report explores explanations for the main finding. The final section of the report considers the implications of the study.

2. Research design and methods

2.1 Intermediate Phase Catch-Up Programme

In 2011 the Gauteng Primary Language and Mathematics Strategy developed and implemented an Intermediate Phase Catch-Up Programme to remediate the learning gaps in underperforming Grade 4 to 6 classrooms. The catch-up programme contains three key elements, namely scripted lesson plans, provision of high-quality learning materials, and on-site coaching. From a research perspective, the catch-up programme had two advantages. First, the designers wrote extensive documentation of all the components and stages of implementation, and, second, they commissioned a rigorous pre- and posttest evaluation from the outset.

The scripted lesson plans divided the term into 11 weeks, with each week designated a number, e.g. Week 8, and each numerical week was linked to a particular calendar week, e.g. Week 8: Monday, 5 March 2012 – Friday, 9 March 2012. Each calendar week for assessment was specified. These seemingly straightforward weekly plans signalled to teachers that they would need to keep up, and that work assigned for the specific work week would have to be completed by the end of the calendar week, so as to ensure that the learners were prepared for the assessment on the specific designated dates.

The original programme used six different learning resources for the classroom. The first was the printed A4 black-and-white lesson plan guide itself. The second was two A4 learner exercise books for each learner, one to write in during the regular class time, and a second specifically for tests. The guide prescribed that the class exercise book be sent home every day, and that the test book be sent home only at the end of the term. The four listening and teaching posters provided to each class cover four themes: “In the Classroom”, “At the Zoo”, “On the Beach”, and “At the Hospital”. The key learning resource provided to all Intersen (Grade 4 to 7) classrooms was a set of ‘reading’ books, what could best be referred to as graded class readers. The guide lists the book title and the week that they are to be used in. The selected titles were listed on the Gauteng Department of Education-approved book list as Grade 2 and 3 books for English Home Language learners. The use of Foundation Phase readers for Grade 6 and 7 learners was informed by the findings of the research (PIRLS, SACMEQ, and ANA), which suggest that most learners in disadvantaged schools are three or more years behind the appropriate grade level in reading in English.

In addition to the A4 exercise books and the 240 reading books, the teachers received a set of “reading sheets”, sufficient for each learner to have one set. The reading sheets contained “look and say” words that learners were expected to know the meaning of, and commit to memory for the formal assessment. The “look and say” words were derived from the reading books, and constitute the core vocabulary and spelling words for the programme. The “look and say” technique, however, did not dominate the catch-up programme’s systematic reading approach, but formed one of three distinct interconnected components, along with a phonics component and the graded class readers, what teachers called the “thin books”. The last learning-and-teaching resource was a mark book, what the programme called the “Assessment Record Book”. The designers of the catch-up programme prescribed a strict and consistent weekly teaching routine, to be followed in the same sequence every week. The teaching week was divided into seven half-hour teaching periods. The

teaching and the homework for each period was specified. Every week was to begin with a "listening and speaking" task, during which teachers teach 10 sentences using the posters. The second period was dedicated to the teaching of phonics and spelling; two new sounds and related words, as well as certain high-frequency words, were introduced. Period 3 was devoted to teaching the "look and say" words that would appear in the class reader for that week. During the fourth period, the teachers were expected to begin using the class reader assigned for the week. The tasks for the period included reading aloud, shared reading, and an oral comprehension exercise based on the class reader. Period 5 was used for consolidation, the sixth period for reading and writing. The final period of the week had two main activities, namely writing and assessment. The assessment took the same form every week, namely a spelling test and a comprehension task. For each period, the guide specified the required homework. Save for the week during which there was to be formal assessment, each week would follow exactly the same format, as the teacher worked systematically through the 12 graded class readers, the four posters, and 12 "look and say" worksheets.

The daily lesson plan guide provided a comprehensive description of each of the 70 lesson periods. A typical daily lesson plan began with a heading which specified the week number, the day of the week, and the date. The lesson time (number of minutes), the lesson outcomes, and the lesson resources were all shown at the top of the page. The 30-minute lessons have either one or two activities. The bulk of the daily lesson plans consists of descriptions of these activities. The activities provide fairly detailed tasks per activity. The lesson plan specified the questions that teachers must write on the chalkboard, and it provided the answers (but it told the teachers not to write these on the board). The 10 questions on the graded class reader vary. Some were simple recall questions from the text (e.g. "Name the fruits that they use to make the fruit salad"), others required the learner's own response (e.g. "Which is your favourite?"), while a few required slightly higher-order engagement (e.g. "Why do they add sugar to the fruit salad?").

The scripted lesson plans and the high-quality learning-and-teaching resources are regarded as a necessary but not sufficient condition for instructional change at scale in this model. The other components, namely the just-in-time training and the ongoing in-class coaching, are viewed as pivotal in changing habits and routines of daily teaching practice. The deployment of instructional coaches was an essential ingredient. The coaches played a number of roles in the programme. They provided training to teachers in small groups, they visited classrooms to model teaching practice and to observe, support, and encourage teachers as they work on the lesson plans, and they monitored and tracked compliance. In the original programme, all coaches were themselves trained in the use of prescriptive protocols for coaching practice.

The theory of change

How do whole-class remedial programmes consisting of the scripted lesson plans, prescribed learner resources, just-in-time training, and in-class coaching change instructional practices and improve learning outcomes? The theory of change embedded in the intervention assumes that these types of interventions, when they are tightly aligned, act to disrupt and re-engineer three core elements of practice. First, the lesson plans and the coaching change how time is understood and used. The first page of the lesson plan guidelines clearly links particular lessons to specific calendar days, thus specifying the pace at which the learning programme is to unfold. The pace remains the same even if teachers are absent or the day is interrupted for any reason. The responsibility, or burden, shifts to

the teacher to keep up with the pre-specified time frames. Within the lesson, teachers need to increase their stamina to keep pace with the relentless forward motion of the lesson plans. The role of the coaches is to assist teachers, and once trust is established, to push them harder to remain on track and to keep up. What the new use of time does is both to increase the amount of time spent on learning tasks and to intensify work on the tasks, thus allowing for increased opportunities to learn and effective coverage of the curriculum. The prescribed weekly lesson routine provided a defined structure to school and lesson time. It is the routine and rhythms of that structure that would enable teachers to cope with the increased pace.

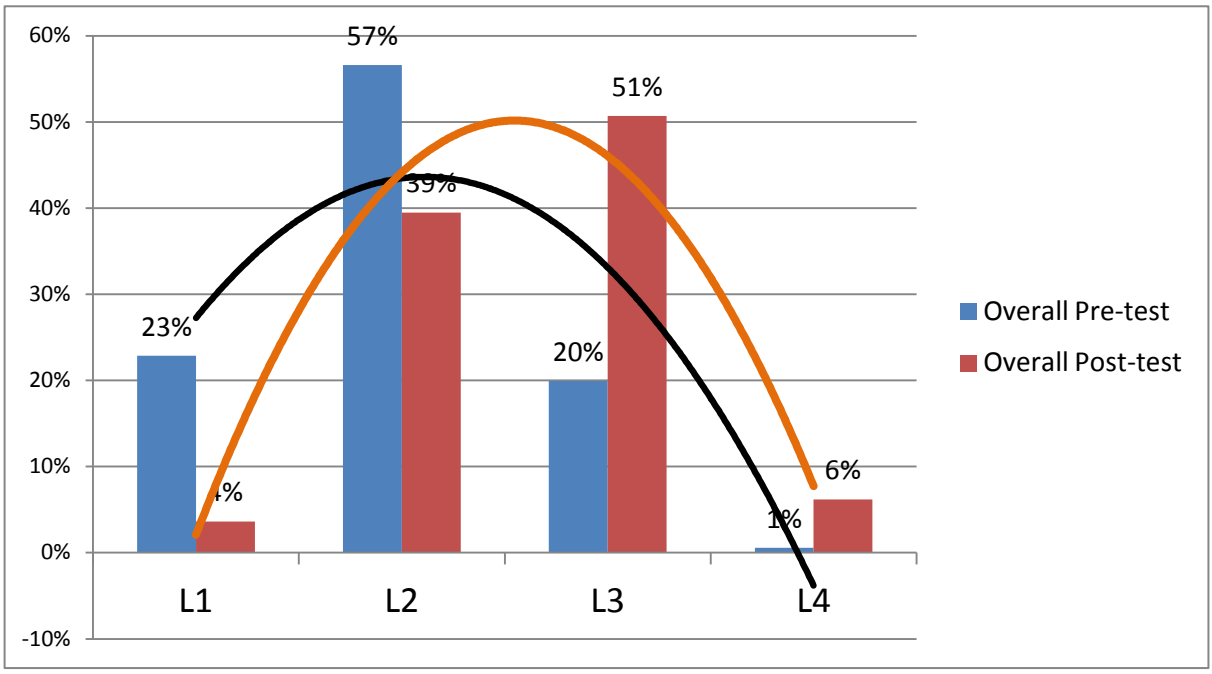
Second, the lesson plans and the learning resources, complemented by the work of the coaches, expand the teachers' pedagogic techniques and classroom management repertoire. One of the consistent findings in the literature (Fleisch, 2008; Carnoy, 2012; Taylor, 2012) is the narrow range of activities and tasks that teachers tend to use. The catch-up programme lesson plans mandate a range of instructional methods and techniques. These included vocabulary development using the wall chart, graded reading using self-contained single-theme readers, systematic phonics, "look and say" word lists, and writing and comprehension strategies. While teachers may have made use of some, or even all, of the methods or techniques at one time or another, the lesson plans provide a systematic and integrated framework within which each method or technique is deployed sequentially and developmentally over time through the carefully structured framework. Teachers experienced not only how the learning tasks embedded in each lesson built on each other, but also how the various methods and techniques, e.g. phonics and class reading, reinforced the learning pathway. The lesson plans also provide tangible instruction on the organisation of time and resources, and classroom management.

The third way that the lesson plans and the learning resources disrupt and re-engineer practice is that they link instruction more directly to the reading levels of most of the learners in the class. An emerging finding in international literature on large-scale reform is the negative consequences of the overambitious curriculum (Pritchett, 2012). By beginning with the average actual reading levels of learners, and moving them systematically along, the intervention ensures that a large proportion of learners will be able to benefit from reading instruction and reading materials at the appropriate grade level by the end of the intervention.

Results of the 2012 catch-up programme pre- and posttest study

The preliminary pretest/posttest evaluation suggested that the programme is effective. The internal study focused on learner performance, and it assessed the extent to which the catch-up programme had improved four distinct literacy skills, namely spelling, language, comprehension, and writing. Two assessment tools were developed, one for learners in Grades 4 and 5, and a second for learners in Grades 6 and 7. The final sample consisted of 1,570 classes, 45% of English Intersen teachers in the project. Hellman (2012) found that, while not all learners were on the same level of achievement at the start of the intervention, across the various different skills, NGOs and districts concerned, the gains made by learners were more or less of the same magnitude.. Overall, a salient outcome of the programme is that irrespective of the grade, NGO, and district concerned, the programme seemed to have had a strong, positive, and consistent effect.

Figure 1: Catch-up programme distribution across the four levels, 2012



Source: Hellman (2012)

While the average scores provide important findings about the magnitude of improvement for the sample, an analysis of the distribution patterns provides additional insights about the effectiveness of the programme.

2.3 Experimental design

The core question that animated this study involved the efficacy and cost-effectiveness of the catch-up programme in improving learner performance on four components of reading. At a theoretical level, the study has the potential to contribute to an understanding of the effectiveness of combining scripted lesson plans, high-quality materials, and instructional coaching.

Until recently, RCT studies were uncommon in developing country contexts.[3] While the findings of these randomised experiments are clearly important, given the high-stakes consequences of their findings, it is necessary to expand the number of studies using these approaches, and to compare findings. One of the problems with some of the existing studies in South Africa is that the evaluations have often been undertaken by the programme developers, which potentially compromises the independence of the investigations.

Randomised control trials have the following features:

- Applying a 'treatment' to the intervention group (e.g. a catch-up programme), while the control group receives either the standard approach or an alternative set of materials or method of instruction.

- Relatively large samples, in order to be able to observe the programme impact with a sufficient degree of statistical certainty (in the case of schools, for example, around 40 or more members in both the treatment and the control groups).
- Randomly assigning schools to intervention and control groups. This is the key step to ensure that all the other factors which influence learner performance should not be any different between the treatment and the control group. Consequently, if there is any observed difference between the groups after the implementation of the intervention, this difference can be attributed to the causal impact of the intervention. This step has earned the method the term "randomised control trial", which is the standard method used in medical research and many other fields of enquiry. Application of the method in the field of education has been fairly recent, although there is now a rapidly growing literature reporting on the use of this approach in the US, as well as on randomised experiments that have been conducted in schools in developing countries, notably in India, Pakistan, Kenya, Uganda, and parts of Latin America.
- Administering pre- and posttests to both groups, and calculating the relative gain score of the intervention group.
- Any statistically significant gains registered by the intervention group can then confidently be attributed to the 'treatment'.

2.4 Site description

The Pinetown district of KwaZulu-Natal Province is the primary research site for the study. It has the advantage of containing a range of poor schools of different types (rural, urban, informal, and formal). It is also conveniently located close to the urban hub. The funder is currently engaged in a larger intervention aimed at improving school primary language and mathematics in the district. As such, the Reading Catch-up Programme may provide useful insights, and if the study shows strong positive results, the programme will be rolled out to all schools in the district.

2.5 Sampling frame and rationale

A detailed report on the sampling procedure is available online in a pre-analysis plan on the RCT registry of the American Economic Association (https://www.socialscienceregistry.org/trials/405). Particular care was taken in designing the most appropriate sampling frame and sample size for the study, to ensure optimal statistical power, as well as to satisfy ethical and cost concerns. As the intervention is designed to improve English reading achievement in underperforming primary schools, we selected only those primary schools where English is the language of learning and teaching (LOLT) from Grade 4 onwards. The second criterion is that only schools that scored 55% or below on the Grade 4 First Additional Language (FAL) test in both the 2012 and 2013 ANA tests in the Pinetown district were eligible for inclusion. The third criterion is that selected schools must have entered between 15 and 120 learners in the FAL Grade 4 ANA test in 2013 (in practice, this number was much higher). This was justified on the grounds of cost. One of the two biggest cost

drivers in this intervention is learner support materials, particularly the graded readers, which are determined by learner numbers and coaches. It is expensive to provide coaching services to schools with fewer than 15 learners in Grade 4. We also excluded schools classified as Quintile 5 schools, which is the most affluent category of schools, according to the official school poverty classification system. Using these criteria, we selected 100 schools to qualify for participation in the study.[4]

For ethical and practical reasons, we sampled intact classrooms within the treatment and the control schools. In other words, all learners in a particular grade in a selected school were included in the study. The ethical reason for doing this is that sampling classrooms within schools would mean that some schoolchildren would receive the benefits of the treatment or control within a single school and grade, and others would not. The practical reason was that if the study had a sub-sample for the treatment or the control within a school, the language teacher would have to be required to teach two different methods simultaneously, which would substantially add to the workload. We assumed, possibly incorrectly, that given the size of the province, and the relative isolation of many rural schools, there would be little danger of a spillover effect from the treatment to the control schools.

With a random assignment to intervention schools and control schools, the variance estimates are large, because schools were the unit of interest, rather than classrooms (if classrooms were the unit of analysis, individual schools could have more than one unit). One of the vexing questions that the researchers grappled with was the number of schools required to ensure that the study could have adequate statistical power.

The study team made the following assumptions:

1. A single language teacher for all Grade 4 classes in each school.
2. Only schools that performed below 55% on the FAL Language 2013 ANA are included.
3. Only schools with between 15 and 120 learners (based on the 2013 ANA) are included.
4. Only public ordinary schools are included.
5. An 80% power level, and a 5% significance level.[5]
6. Testing restricted to a random sample within a single grade.
7. ICC value (between-school variance as a proportion of total variance) of 0.20.[6]
8. Oversampling of control schools relative to intervention schools.
9. A correlation between pretests and posttests of 0.7.
10. Attrition among learners would not pose problems to the integrity of the study. Since the pre and post testing occurs within a 12-week period, absenteeism was probably going to be the main cause of attrition, and this would not likely to be systematically different between treatment and control groups. Consequently, attrition would not bias the estimated treatment effect.
11. Minimum detectable effects (MDE) set at 0.2 standard deviations.[7]

Given these assumptions, a sample size of 40 treatment schools and 60 control schools was adequate. A computerised lottery was used to randomly allocate schools in the final sampling frame into the treatment and the control groups.

Ultimately, these sampling assumptions proved to be conservative – a particularly low intra-class correlation coefficient (0.15) and a high correlation between baseline test scores and endline test scores (0.8) meant that the study was actually powered to identify a minimum detectable effect size of 0.15 standard deviations, which turned out to be about 3.5 percentage points in the reading test. This means that if the true impact of the intervention was to improve reading test scores by 3.5

percentage points (relative to the control group), then we will be 80% sure to obtain a statistically significant estimate of the treatment effect.

In addition to measuring the short-term effect of the intervention on average grade reading performance, we also planned on using official data from Annual National Assessments to measure the longer-run impact of the programme on language achievement. This would provide important evidence on the extent to which short-term remedial interventions, such as the catch-up programme, can lead to improvements in educational outcomes.

2.6 Pretest learner results

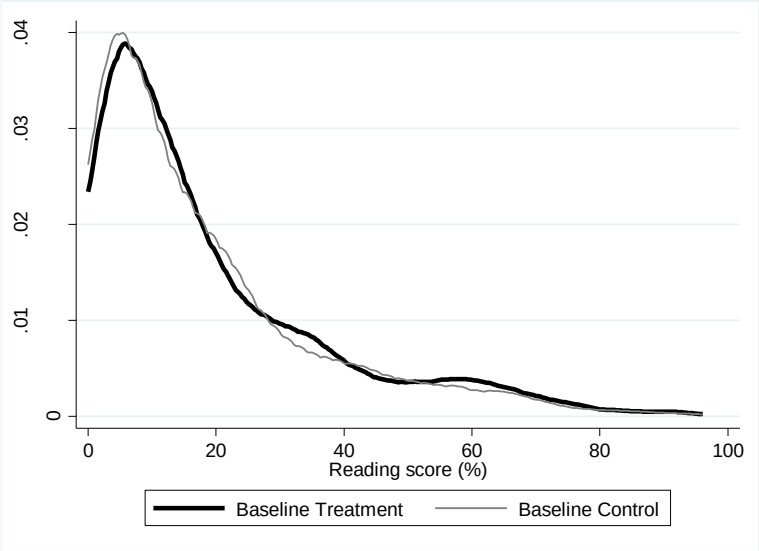
This subsection begins with descriptive information on the intervention in Pinetown. This is followed by a presentation of key data from the pretest.

The original intention was to have a balance of 40 treatment schools and 60 schools in the control group. One problem that occurred was the need to replace three control schools just before the pretesting began. These schools were replaced at the request of the district office, and the reasons provided were legitimate, and would have applied equally to treatment schools, had it been necessary. This meant that the remaining 57 control schools still serve as a valid comparison group to the treatment schools. For the calculation of results we thus used only these 57 control schools, and did not use the three new control schools, because these were non-randomly added by the district office, therefore potentially compromising the validity of the control group. A further challenge was that one control school did not participate in the baseline testing, but did participate in the endline testing. We therefore did not have baseline data for this school.

We obtained data on the pretest for 2,663 learners from 96 schools. For purposes of analysis, however, we only used data from the 2,543 learners who also wrote the posttest. The focus of the data analysis of the pretest was on the effectiveness of test items, and to check the balance between the treatment schools and the control schools.

There were 36 numbered test items, and a few items with multiple components. As such, the maximum possible score was 51. The first analysis was designed to ascertain the number of learners with non-responses on items. Non-responses could have been due to no answer provided, or more than one response provided. Seventy-five percent of children had six or fewer items with no response. This was positive. Our plan for calculating test scores was to regard non-response as incorrect. Figure 2 shows the distribution of baseline scores (expressed as percentage scores) for learners in both treatment schools and control schools. The figure indicates how similar the distributions of achievement were between treatment and control schools, confirming that the randomisation was successful in producing adequate balance between the two groups. Figure 2 also shows that the vast majority of the learners scored below 20% on the pretest. Given the very low scores on the pretest, concerns were raised about a possible ‘floor effect’. This may have had the unfortunate effect of making it harder to identify improvements in learning at the bottom end of the distribution.

Figure 2: Kernel density of pretest scores, percentage



The questions on the cover of the test instrument allowed the research team to analyse some of the characteristics of the study population. Tables 1 and 2 and figure 3 show the performance averages and distributions by age and gender.

Table 1: Baseline performance by age of learners

Age	Mean reading score	Number of learners
8	27.09	11
9	21.80	1,072
10	17.41	832
11	13.56	324
12	10.69	97
13	9.59	46
14 and older	17.29	148
Age not specified	16.89	13
Average/Total	18.41	2,543

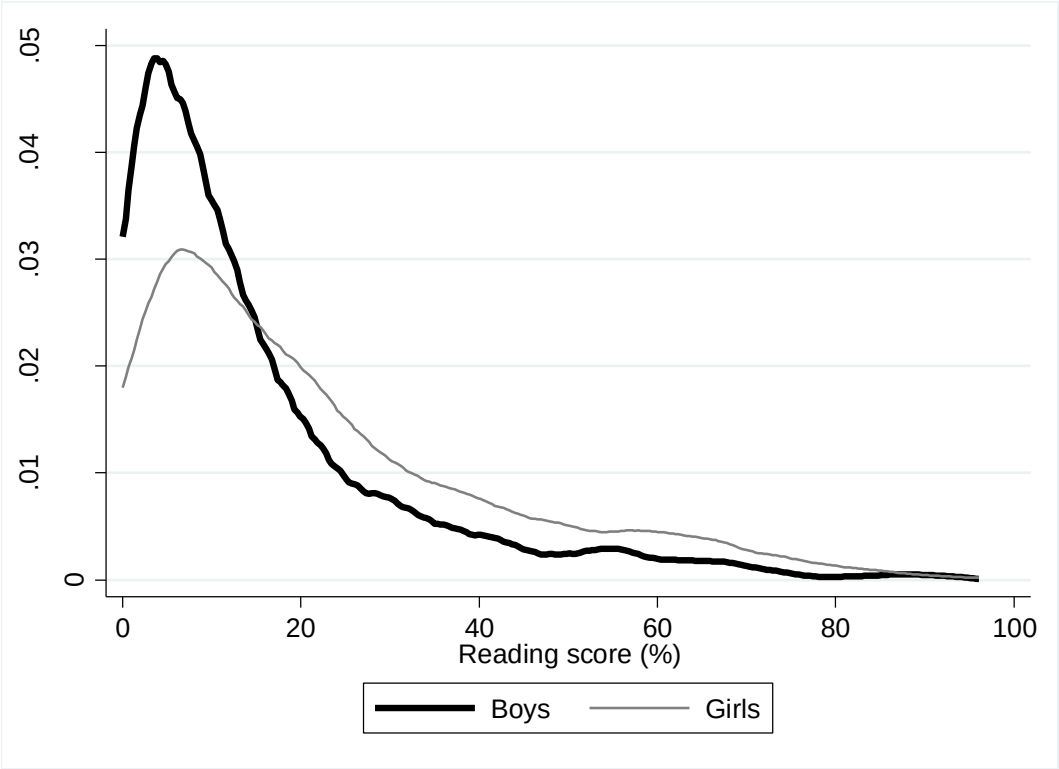
Table 1 reveals that, on average, schoolchildren at the 'correct' age to grade had the highest mean scores, with the scores dropping substantially for older learners. What is of concern is the relatively large number of learners (148 out of 2,543) that reported their age as 14 years or older, which is five full years beyond the norm for Grade 4.

Table 2: Baseline performance by gender

Gender	Mean reading score
Boys	14.89
Girls	22.05
Average	18.40

Table 2 and figure 3 reveal the gender imbalance in performance, with girls substantially outperforming boys in the overall sample. This is in line with other test results from South Africa, such as PIRLS 2011 and the Annual National Assessments of recent years, which all show a significant test score advantage for girls, particularly in literacy.

Figure 3: Distribution of baseline performance by gender



2.7 Implementation

The information on the implementation of the treatment has been drawn from the service providers' monitoring report. The implementing agency classified the school into three groups – green, yellow, or red – on the basis of their implementation after week 3, based on the coaches' assessment of their commitment level and teacher competency. These ratings are subjective.

Table 3: Level of implementation and teacher competency

	Total number of teachers	Percentage
RED	7	11
YELLOW	23	34
GREEN	37	55

Source: Class Act Monitoring Report

The implementing agency reported the following challenges:

- Teachers felt that the pace required by the project was too fast, and they were not used to preparing for or implementing 10 English lessons per week, despite CAPS requirements.

- The second major challenge was related to compliance: preparation, planning, and implementation. The afternoon workshops addressed this challenge to some extent, but teachers that did not attend did not get the benefit of these planning sessions. The response to this was initially to offer additional support to non-compliant teachers. However, from mid-May a decision was made to focus coaching attention on more committed teachers. Non-compliant teachers and principals were aware that posttesting would be implemented.
- Teachers needed support with the technical process of working out averages. In response to this, the implementing agency introduced a 'reward system'. Once teachers were up to date with submissions, and the submissions had been verified against learners' books, teachers received a pack of stamps/stickers to use when marking the learners' books.
- The poor quality of written work was identified as an ongoing challenge. Teachers generally gave poor instructions, and did not give enough support with regard to written work.
- The management and use of classroom resources and the environment was another challenge. Teachers did not display the flashcards and other resources in a meaningful way, to reinforce learning that had taken place.
- The use of code-switching was pervasive. Some teachers taught the entire English lesson in isiZulu, using English only for key words or phrases.
- Most teachers appeared to welcome the structure, routines, standardised methodologies, and content of this project. There was some evidence of improved time on task and work rate, despite the constant tension around pacing.

Table 4: Curriculum coverage

	Phonics	Listening and speaking			Shared reading and oral comprehension		Spelling Tests		Written comprehension		Grammar and writing	
	Total # lessons	Lessons completed	Total # lessons	Lessons completed	Total # lessons	Lessons completed	Total # lessons	Lessons completed	Total # lessons	Lessons completed	Total # lessons	Lessons completed
Average number of lessons completed	21	14	13	9	19	12	8	6	8	5	10	8
Average percentage of lessons completed		74%		69%		63%		67%		63%		70%

Table 4 provides important information about curriculum coverage. On average, teachers enacted between 63% and 74% of the planned lessons. Teachers covered the phonics lessons more completely than they did the shared reading and oral comprehension and the written comprehension lessons. Although two-thirds of the lessons were far from a desirable level of programme completeness, the programme completeness is certainly higher than most estimates of the proportion of the curriculum that teachers tend to cover.

3. Results

From the perspective of the study design, one of the most positive outcomes of the posttest was the low level of attrition observed between the pretest and the posttest. No entire schools were lost on follow-up. Table 5 shows that attrition among learners appears to have been low, and that it was not particularly skewed across treatment and control groups.

Table 5: Attrition between pretest and posttest, RCUP 2014

	Present at endline	Not present at endline	Total
Control	1,423	127	1,550
	(91.81%)	(8.19%)	(100%)
Treatment	1,043	70	1,113
	(93.71%)	(6.29%)	(100%)
Total	2,466	197	2,663
	(92.6%)	(7.4%)	(100%)

Overall, of the 2,663 learners that wrote the pretest, 2,466 completed the posttest, which represents a 7.4% attrition rate. The attrition rate was slightly higher in the control group than in the treatment group. When running a regression to test whether allocation to the treatment group predicts attrition, it is evident that treatment does not predict attrition at all once variables such as baseline scores are controlled for. Therefore, we excluded learners that were absent from the dataset, and proceeded to analyse the data using only learners that were present in both the pretest and the posttest.

The core question that animated this study focuses on the extent to which learners' achievement in English literacy improved as a result of exposure to the Reading Catch-Up Programme. The data show only a very small difference in posttest means between control and treatment school groups.[8] A comparison of the trend lines in the pre- and posttests for the treatment and control schools shows that while both groups improved substantially between the pretest and the posttest, the improvement is only marginally better in the treatment group. In other words, while the baseline trends were very similar, so were the endline trends.

Figure 4: Posttest score distributions for treatment and control schools

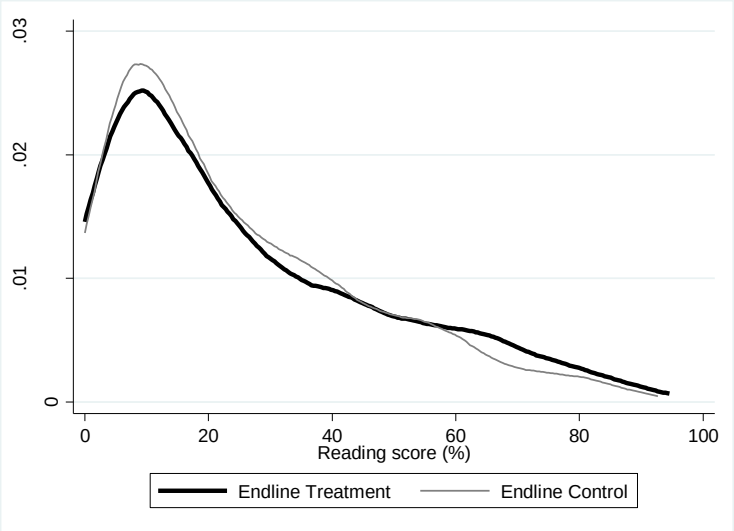
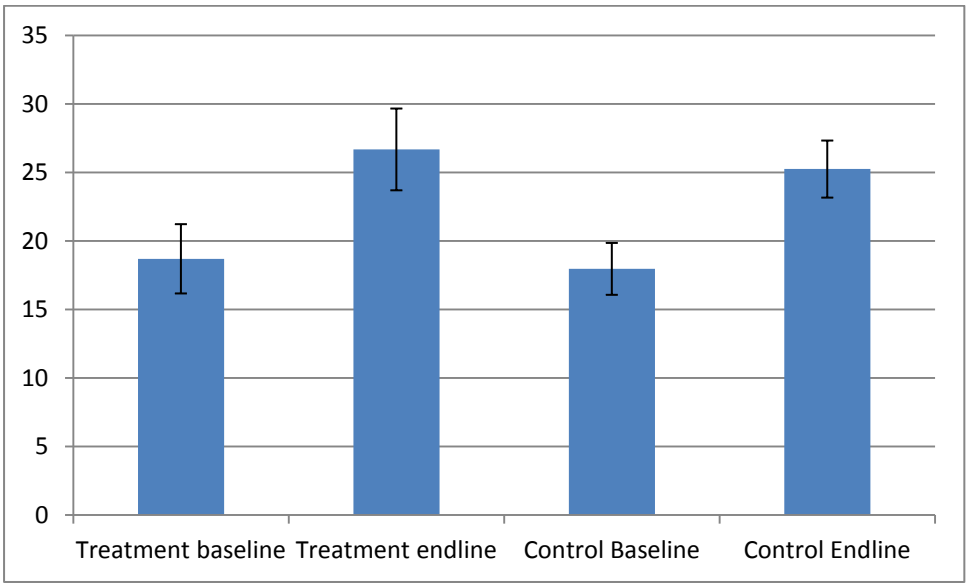


Figure 5: Mean scores for treatment and control groups (pre- and posttest)



Note: 95% Confidence Intervals are indicated

The small difference in improvement in the treatment schools relative to the control schools is clear upon observation of figure 5. In statistical terms, although the posttest score was higher in the treatment schools than in the control group, the difference is not statistically significant.

Table 6 shows the results of five regression models, which represent the most robust methods for estimating the impact of the programme. Column 1 represents the model where the outcome variable is the overall score on the posttest, or endline literacy test. The main explanatory variable of interest is a variable indicating whether the school is a treatment school or a control school. Other variables included in the regression model are the learner's baseline, or pretest score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Although there

is no reason to expect differences in endline test score between the treatment schools and the control schools *as an effect of causes other than the intervention*, it is still worth including these other control variables, in order to enhance the statistical precision of the estimated treatment effect. Only the coefficient on the treatment variable and the standard error of the estimate are reported in Table 6, but all the above-mentioned controls were included. Columns (2)-(5) in the table represent models with the same set of explanatory variables, with the difference being that the outcome variables are learner scores for each of the four literacy domains which formed part of the reading test.

Table 6: Main regression results

	(1)	(2)	(3)	(4)	(5)
	Overall score	Spelling	Language	Comprehension	Writing
Treatment	0.49	1.27**	3.96***	-1.40	1.14
Standard error	(0.67)	(0.61)	(1.07)	(1.34)	(1.40)
Observations	2,466	2,466	2,466	2,466	2,466
R-squared	0.77	0.77	0.46	0.53	0.28

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

The estimated treatment effect on the overall literacy score is 0.49 percentage points gained relative to the control group. However, we are unable to conclude with any level of statistical confidence that the true effect is statistically significantly different from 0. On the other hand, we are able to conclude with high levels of statistical confidence that the intervention improved spelling outcomes and language outcomes for learners in treatment schools. We estimate that spelling improved by 1.27 percentage points relative to the control group, and that language improved by 3.96 percentage points. The estimated impact on comprehension and writing items was not statistically significantly different from 0.

Heterogeneous treatment effects

We also investigated so-called "heterogeneous effects", that is, whether the impact of the programme was different depending on various learner, school or teacher characteristics. There was no evidence of heterogeneous effects based on learner gender, learner age, learner exposure to English at home, or class size (full results are not reported here). In planned forthcoming analyses we will continue to investigate heterogeneous effects according to other characteristics, as outlined in the pre-analysis plan.

The following analysis (see table 7), however, points to the possibility that the impact was larger for children that initially performed better on the baseline test. The result is statistically significant for spelling. Although the size of the coefficient for language is not statistically significant, it is actually

larger than that of the coefficient for spelling, so it may be that the same was true for language, but we are simply unable to conclude so with statistical confidence. For spelling, there was effectively no impact on those learners that had initially scored poorly (and there were indeed many scores of 0). The coefficient on the interaction term indicates that every 10 percentage points gained relative to the baseline test was associated with an increased treatment effect of 0.5 percentage points.

Table 7: Impact by baseline performance of learners

	Combined score	Spelling	Language	Comprehension	Writing
Treatment	-0.44	0.32	2.92**	-1.96	1.39
	(0.86)	(0.7)	(1.15)	(1.54)	(1.47)
Baseline percentage score	0.97	0.93	0.66	0.79	0.49
	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Treatment **X** baseline	0.05	0.05*	0.07	0.02	-0.03
	(0.04)	(0.03)	(0.04)	(0.04)	(0.06)
N	2,466	2,466	2,466	2,466	2,466
r2	0.77	0.77	0.46	0.53	0.28

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

Effects based on differing treatment intensity

It is possible that the benefits of the programme would have differed depending on how enthusiastically teachers participated. The service provider collected information that serves as a proxy for level of commitment, that is, the number of lessons completed (see table 4). Altogether, 79 lesson plans could have been implemented over the period of the intervention, and, on average, teachers completed 66% of the lessons. Five teachers completed less than half of the lessons. Eight teachers completed at least 75% of the lessons. We also have information on the number of class assessments that each teacher completed (these were provided as part of the intervention programme). Twenty-seven teachers completed 12 assessments, and 13 teachers completed fewer than 12 assessments. The service provider also recorded attendance at afternoon workshops held in small clusters in intervention schools. As Table 8 indicates, some teachers attended only one or two afternoon workshops, while others attended five or six.

Table 8: Teacher attendance at afternoon workshops

Number of training sessions attended	Number of teachers
1	9
2	16
3	11
4	8
5	4
6	7
Total	55

The main estimate of the programme impact, as reported in table 4 is conventionally referred to as the "intention-to-treat" (ITT) estimate, where allocation into the treatment group indicates an intention that these schools receive the intervention. However, when compliance with the intervention is not uniform, we are also interested to measure what is called the "treatment-on-the-treated effect" (TTE), that is, the effect of the intervention for those who complied with the intervention. Recovery of a TTE estimate is more complicated statistically than the simple ITT, since we cannot identify a valid comparison group of teachers in the control group who would have complied had they been in the intervention group. The appropriate statistical model is to use treatment allocation as an instrumental variable (IV) for the level of compliance observed. For example, we can make the explanatory variable of interest the number of RCUP lessons completed by teachers (which is always 0 for control schools), and then use treatment allocation as an instrumental variable. We use two-stage least squares regression analysis to obtain the estimates of the TTE, as reported in Table 9.[9]

In Table 9 we report TTE estimates for the main outcome (reading test total score) and the two subcomponents of the test for which we observed a statistically significant positive impact (spelling, and language). For the purposes of the analysis, we decided to specify high compliance as a binary variable, taking a value of 1 if compliance was high, and 0 if compliance was low. In the case of assessments completed, it made sense to define high compliance as having completed 12 of the assessments provided, since the majority of teachers had completed exactly 12 assessments. We use three different options for defining high lesson completion: definition 1: completed at least 50% of lessons; definition 2: completed at least 60% of lessons; definition 3: completed at least 70% of lessons. Similarly, in the case of training workshops attended, we compare results when using three alternative definitions of high attendance: definition 1: attended at least three sessions; definition 2: attended at least four sessions; definition 3: attended at least five sessions.

In all models in Table 9 the basic result is similar: there is no significant effect on the overall reading score, but there are significant positive effects observed for spelling and language. The TTE estimates are all larger than the corresponding ITT estimates, as shown in Table 7. This indicates that when teachers complied more with the programme, the impact was indeed larger. One caveat to this conclusion is that although the TTE for spelling and language were both statistically significantly different from 0, they were not statistically significantly larger than the ITT estimates. Nevertheless, the fact that in all regressions the TTE estimate was larger than the ITT estimate means that there is suggestive evidence of a greater benefit to learners when teachers implemented the programme extensively. The size of the difference is important, but we would have needed to have a larger sample to achieve more precise estimates. This analysis points to the possibility that the success of an intervention such as the RCUP may depend on the extent to which teachers engage with it.

Table 9: IV estimates of the treatment effect on the treated

	Overall score			Spelling			Language		
	Def 1	Def 2	Def 3	Def 1	Def 2	Def 3	Def 1	Def 2	Def 3
Lessons completed	1.02	1.14	1.46	2.16***	2.41***	3.10***	5.30***	5.91***	7.60***
(IV: treatment)	(-0.93)	(1.05)	(1.34)	(0.81)	(0.92)	(1.20)	(1.58)	(1.82)	(2.27)
Training workshops	1.12	1.78	2.72	2.40***	3.78**	5.79**	5.82***	9.26***	14.15***
(IV: treatment)	(1.03)	(1.72)	(2.73)	(0.89)	(1.56)	(2.65)	(1.75)	(3.23)	(5.56)
Assessments	1.50	-	-	3.19***	-	-	7.83***	-	-
(IV: treatment)	(1.38)	-	-	(1.20)	-	-	(2.43)	-	-

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

Did the impact of the intervention depend on which coach the school was allocated? The service provider used two coaches to implement the programme. Each coach was allocated 20 schools. Therefore, one can estimate two separate treatment effects, one for each coach. Table 10 shows the results when running the exact same regression models as reported above, but instead of including a single treatment dummy variable, we include two dummy variables (one for each coach), still relative to the reference category of control schools. There are two main limitations in this analysis. First, the coaches were not randomly assigned to schools. However, the fact that we have baseline scores for each learner, and can control for stratification and other learner, school and teacher characteristics, reduces the likelihood of omitted variables bias. Second, the effective sample size is cut in half: instead of having a treatment group of 40 schools, we now compare each treatment group of 20 schools to each other, and to the control group. This means that standard errors will be larger, and that therefore we are less likely to observe a statistically significant treatment effect.

Table 10 shows no significant impact for coach B on any of the outcomes. For coach A, however, there were statistically significant effects on both spelling and language. The coefficients for coach A are all larger than in the overall treatment effects as reported in table 5 (though we cannot conclude with statistical certainty that the effects are larger). Therefore, this provides suggestive evidence

that the success of an intervention that uses coaches to support teachers may depend on the particular person doing the coaching. If indeed this was the case, we are not able to determine what characteristics of coach A led to a larger impact.

Table 10: The impact of the coaches

	Combined score	Spelling	Language
Coach A	1.42	1.98**	5.87***
	(0.93)	(0.76)	(1.42)
Coach B	-0.42	0.56	2.09
	(0.89)	(0.88)	(1.41)
N	2,466	2,466	2,466
r2	0.7698	0.7692	0.4606

Note: *p<0.1 **p<0.05 ***p<0.01

The reference category for both coaches is the control group. All models include controls for baseline score, stratification dummies, learner gender, learner age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

Impact on Annual National Assessments

The Annual National Assessments (ANA) of 2014 were written during the week of 16-19 September across schools in South Africa. This was about three months after the RCUP intervention was completed. All children in Grades 1 to 6 and 9 wrote a mathematics test. Children in Grades 1 to 3 wrote a Home Language test (in Pinetown it was in isiZulu). Children in Grades 4 to 6 and 9 wrote one of the following language subjects: English Home Language, Afrikaans Home Language, English First Additional Language, or Afrikaans First Additional Language. In Pinetown, 94% of learners in our sample of treatment and control schools wrote English as First Additional Language.

There are several hypotheses which the availability of ANA data allowed us to investigate:

i. *The treatment effect for intervention schools relative to control schools may diminish over time, or it may grow through continued use of the new materials and pedagogies.*

ii. *An improvement in literacy may benefit other learning areas, such as mathematics.*

iii. *Although the intervention targeted Grade 4 teachers in a school, there may be spillover benefits to other grades.*

The third hypothesis is especially possible since the majority of Grade 4 teachers in South Africa also teach in another grade.[10] We used ANA data for literacy in Grades 1, 2, 3, 5 and 6, to see whether students in untreated grades in intervention schools improved relative to students in control

schools. We also used ANA data for mathematics in Grade 4, to ascertain possible impact of the treatment on other subjects.

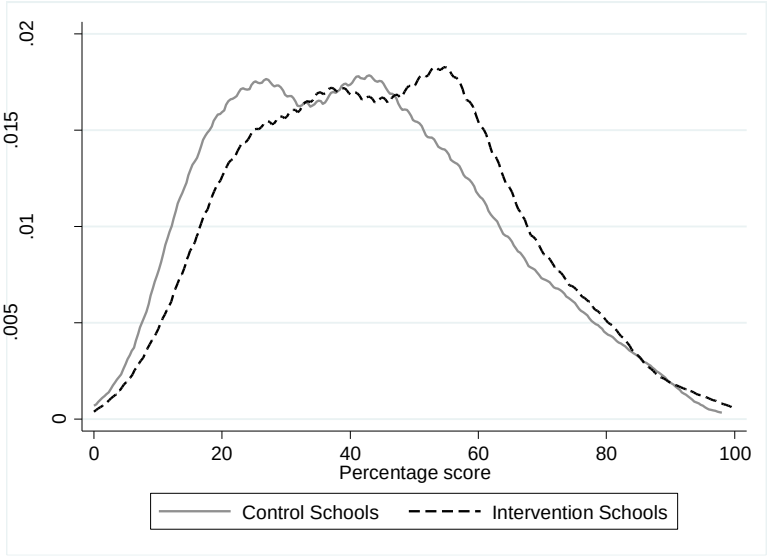
There are, however, several limitations of the ANA data for our purposes. The data quality is not expected to be as high as that collected by our service provider. This is because the ANA tests were locally administered and marked by teachers within each school. Differences in the conditions of testing, and in marking standards across schools, should make the data a somewhat noisy signal of learner proficiency. This is confirmed by the respective correlations between our baseline test score, our endline test score, and the ANA language scores of learners. In a sample of 1,928 learners who we were able to match between the RCUP and ANA datasets, the correlation coefficient between the baseline test score and the endline test score was 0.86. However, the correlation coefficient between baseline score and ANA English score was only 0.53, and between endline score and ANA English score it was 0.56. Noisy data would be expected to cause a degree of attenuation bias in the estimated treatment effects (where the estimated effect is biased towards 0). Fortunately, though, there is no reason to expect differences in marking, or quality of ANA information, to be correlated with assignment to treatment.

In the first analysis using ANA data, we use all learners in treatment and control schools, that is, not only those learners that were sampled for our own independent testing. This provides us with a dataset of 6,419 learners across our treatment and control schools. While this improves the statistical power to identify a treatment effect, the disadvantage of this approach is that we do not have a baseline score for each learner. The best we can do is to control for each school's average ANA score in previous years.

The average score in Grade 4 English First Additional Language within our sample of schools was 43.0%. As was the case in our independently administered tests, the girls (average score of 46.8%) substantially outperformed the boys (average score of 39.4%). Importantly, the male disadvantage was still large (about 6 percentage points) in all our multivariate regression models, even after controlling for other characteristics, such as age (the boys are noticeably older than the girls, on average). Although this finding is not central to this report, it confirms an increasingly clear pattern of a large learning disadvantage for males in schools in South Africa.

Figure 6 presents kernel density curves which show the distributions of test scores for learners in intervention schools and learners in control schools. This indicates that learners in intervention schools had a somewhat better distribution of achievement than did learners in control schools. This is a preliminary indication of a positive treatment effect.

Figure 6: Kernel **d**ensity **c**urves of **t**est **s**cores for Grade 4 English First Additional Language (ANA, 2014)



The first hypothesis to test is whether learners in intervention schools performed better in the Grade 4 English ANA test than did learners in control schools. When no attempt is made to control for baseline differences in achievement, the estimated treatment effect is 3.35 percentage points, and this is statistically significant at the 90% level.[11] Models 2, 3, 4 and 5 in Table 10 show the estimated treatment effect when different ways of controlling for prior school performance are used (namely controlling for school mean language score in ANA 2013, controlling for school mean language score in ANA 2012, and controlling for school mean language score in both ANA 2012 and ANA 2013). In all cases, the estimated treatment effect is somewhere between 3 and 4 percentage points, but in models 4 and 5 it is not statistically significant.

Table 11: **Treatment effect on Grade 4 English First Additional Language (ANA)**

	Model 1	Model 2	Model 3	Model 4	Model 5
Treatment	3.35*	3.83**	3.49**	3.13	3.26
	(1.93)	(1.88)	(1.72)	(1.95)	(2.14)
School mean Grade 4 2013	No	Yes	Yes	No	Yes
School mean Grade 4 2012	No	No	Yes	No	No
School mean RCUP baseline	No	No	No	Yes	No
School mean Grade 3 2013	No	No	No	No	Yes
N	6,419	6,419	6,419	6,419	6,055
r2	0.1731	0.1914	0.2072	0.1753	0.2042

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for stratification dummies, learner gender, and learner age. Standard errors are adjusted for the fact that learners are clustered in schools. Since a few schools wrote English as Home Language, only 91 schools are represented in the models (37 treatment schools, and 54 control schools). The results are robust to an alternative specification where the outcome variable is percentage score, irrespective of whether this was from the English as Home Language test or the English as First Additional Language test.

Was there a spillover benefit observed in Mathematics scores of learners that had been exposed to the catch-up programme? Since the Mathematics test is formulated in English, it is plausible that an improved English proficiency thanks to the RCUP intervention would have led to improved Mathematics scores. As reported in Table 11, although the estimated treatment effect on Mathematics scores was positive, it was not statistically significant. Therefore, we cannot conclude that the intervention led to improved Mathematics performance.

Table 12: The effect of treatment on Grade 4 Mathematics (ANA)

	Model 1
Treatment	2.38
	(2.58)
Baseline school average 2013	Yes
Baseline school average 2012	Yes
N	6,687
r2	0.2153

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for stratification dummies, learner gender, and learner age. Standard errors are adjusted for the fact that learners are clustered in schools.

Was there a spillover benefit observed in language performance for other grades at treatment schools? The results in Table 13 indicate that there was a positive effect for the grades on either side of the treated group, that is, Grade 3, and Grade 5. The fact that the majority of Grade 4 teachers in South Africa teach in another grade strengthens the plausibility of this result. On the other hand, it seems less likely that Grade 3 Home Language (isiZulu) would improve through an English intervention in Grade 4. Therefore, we recommend that no strong conclusions be made on the basis of this result.

Table 13: Treatment effect on language across untreated grades (ANA)

	Grade 1	Grade 2	Grade 3	Grade 5	Grade 6
Treatment	1.55	-1.01	5.80***	3.49**	3.26
	(1.82)	(1.50)	(1.80)	(1.60)	(2.07)
Baseline school average 2013	Yes	Yes	Yes	Yes	Yes
Baseline school average 2012	Yes	Yes	Yes	Yes	Yes
N	9,144	7,673	7,089	5,341	4,963
r2	0.1131	0.0958	0.1577	0.2083	0.2407

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for stratification dummies, learner gender, and learner age. Standard errors are adjusted for the fact that learners are clustered in schools. Since a few schools wrote English as Home Language, only 91 schools are represented in the models (37 treatment schools, and 54 control schools). For Grades 1, 2 and 3 the test was a Home Language test, while for Grades 5 and 6 the test was English as a First Additional Language.

As before, we test whether there was a different treatment effect for each coach. The results are very similar to those observed when using the independently administered test data. For coach A there was a fairly large and statistically significant treatment effect, while no significant effect was observed for coach B. However, as before, we cannot actually say with statistical certainty that the effect for coach A was larger than that for coach B.

Table 14: The impact of the coaches (ANA)

	ANA language
Coach A	5.38**
	(2.45)
Coach B	0.85
	(1.70)
N	6,419
r2	0.2106

Note: *p<0.1 **p<0.05 ***p<0.01

The model includes controls for school mean language score in both ANA 2012 and ANA 2013, stratification dummies, learner gender, and learner age. Standard errors are adjusted for the fact that learners are clustered in schools.

Analysis of the sub-sample of individuals that participated in both RCUP testing and ANA

Out of the 2,466 learners with valid pre- and posttest scores in the final RCUP dataset, we were able to identify 1,928 learners in the Universal ANA dataset of 2014. We matched learners by using the first three letters of their first names, the first three letters of their surnames, their gender, their school, and their grade. This led to some duplicates, where matches were identical, based on these variables. We therefore excluded from the sub-sample all individuals affected by such matches, to avoid the possibility of false matches. There are several other possible reasons why we would have

not identified all learners in the ANA data. It may have been that some learners were absent on the day of the ANA testing. Some learners may have participated in the Verification ANA testing, in which case their ANA marks would not be captured in the Universal ANA dataset. Some learners may have participated in Universal ANA, but due to incomplete data capturing, their results were not uploaded onto the national dataset. There may have been errors in the information used to match learners across the two datasets, that is, their name or surname may have been misspelt in one of the datasets.

The advantage of using individuals with both RCUP information and ANA test scores is that we can control for a baseline score for each learner, namely the baseline score on the RCUP test. We ran a regression analysis to check whether treatment status predicts being successfully matched in the ANA data. This indicated no statistically significant relationship between being in a treatment school and being found in the ANA dataset. Therefore, we can analyse the results on the ANA tests for treatment and control schools without fear of any selection bias that might influence the estimated treatment effect. This is further confirmed by the fact that when we run the exact same regression analysis as that used in the main model (that is, predicting RCUP endline scores), as reported in Table 6, but on the sub-sample of 1,928 matched learners, we obtain essentially the same estimated treatment effect (namely a coefficient of 0.48, as opposed to 0.49).

Table 15 reports the results of the two models that we ran on the individually matched sub-sample. The outcome variable is percentage score in Grade 4 English as First Additional Language. The magnitude of the coefficients observed in Table 15 are broadly consistent with results reported thus far in the paper, namely a relatively small positive effect of being in the treatment group, a larger positive effect for coach A, and a negligible effect for coach B. However, all the coefficients of interest in these two models are not statistically significantly different from 0. The effect sizes are non-negligible, which means that we were somewhat underpowered, particularly in the case of the coach-specific models. The overall conclusion to draw from this analysis remains as follows: there is tentative evidence of a fairly small effect of the intervention on performance, and this effect appears to have been larger for coach A, but we cannot make these conclusions with a high level of statistical certainty.

Table 15: The impact of treatment on ANA language scores (for individually matched sample)

	Model 1	Model 2
Treatment	2.40	
	(2.20)	
Coach A		4.89
		(3.39)
Coach B		0.14
		(2.38)
N	1,928	1,928
r2	0.4643	0.4676

Note: *p<0.1 **p<0.05 ***p<0.01

4. Discussion

Even though the increases in learners' spelling and language scores in the treatment schools are statistically significant, and the ANA scores show statistically significant relative gains compared to the control schools, the gains may have limited educational significance. The effect sizes, as measured by standardised scores, were relatively small compared to the gains suggested in the original 2012 reading catch-up study, and in Pretorius' (2014) new study.[12] A scan of some learner posttest scripts sampled from treatment schools clearly shows that most of the Grade 4 learners continue to be very weak spellers, with limited command of basic structures of the language, and poor comprehension and writing proficiency. The gap between these learners' literacy performance and the demands of the curriculum remains large.

The core hypothesis that Intermediate Phase learners' literacy proficiency could be 'caught up' across a 'subsystem', using a well-designed ten-week intervention, is simply not supported by the evidence from this randomised control trial. That said, there is evidence to suggest that with higher levels of implementation intensity and/or extended duration, and with strong coaching, interventions such as the Reading Catch-Up Programme could indeed enable learners to narrow the gap between their actual literacy performance and the expectations of the official curriculum, particularly in domains such as spelling and language. The potential for improvement through this sort of programme would seem to be greater for those learners that are not at the very bottom of the performance distribution.

Explaining the limited gains in the treatment group

Before exploring substantive reasons for the low estimated impact of the programme on reading outcomes, it is worth highlighting a few possible measurement limitations that may have contributed to this outcome. While there was a substantial increase between the pretest and the posttest, the gains were very similar for treatment and control schools. Why would there be such a dramatic gain in the control group? A number of explanations can be offered. First, it may simply be that soon after beginning with English as the language of instruction (as occurs in Grade 4), learners typically demonstrate rapid gains in basic vocabulary. If this is the case, then the large gain in the control group is perfectly legitimate, and in no way biases the results of this study.

Another possibility relates to the Hawthorne effect, that irrespective of whether a school was assigned to the control group or the treatment group, all schools were subject to external scrutiny, particularly around learner performance testing (that is, pre- and posttesting). The very fact of being tested by an external agency, in and of itself, might have been the impetus for more engaged teaching and learning, particularly as schools are increasingly concerned about possible high-stakes consequences of the new annual national testing policy. If a Hawthorne effect was present for the control schools, then this is not a problem for the study design, since treatment schools would also have experienced a Hawthorne effect as a result of being tested, and these effects would cancel each other out. We are precisely interested in the effect of the programme over and above any effects of testing.

A potentially problematic possibility is that there was an unanticipated spillover effect, where schools that were part of the control group received some of the benefits of the RCUP intervention through informal sharing between schools. Further analysis of the data will be conducted to investigate whether this may have occurred, but it seems unlikely that this would have occurred to

any great extent, since the main aspects of the programme (namely coaches and materials) were not easily transferable.[13]

A third explanation may be found in the 'floor effect' evident in the pretest results. While the decision to employ the identical instrument used in the original Gauteng study was deliberate, and would theoretically have allowed for precise comparison of gain scores, the context in KwaZulu-Natal might mean that learners in that province have considerably lower access to English vocabulary and literacy in English in general than is the case for their counterparts in Gauteng. A different instrument, one that emphasised Grade 1 English FAL questions, might have provided results more closely resembling a normal distribution. Such an instrument might have revealed gains at the lowest levels of literacy.

Notwithstanding the above questions, the statistically significant findings of gains in two domains, namely spelling and language (grammar), are important. These are clearly the domains most likely to change, as they have the lowest cognitive load associated with them. Should learners have encountered the spelling words directly during the 10 weeks of lessons or mastered some aspect of English phonics, it would be reasonable to expect that this learning would be evident in the posttest and a few months later in the ANA test. Similarly, explicit teaching of punctuation, such as a capital letter at the beginning of the sentence, and a full stop at the end of a sentence, would carry through to improved scores on the language section of the posttest. In contrast, the fact that scores did not change for comprehension, which requires a much wider and more complex range of knowledge and skills to be taught and learnt, is not surprising, given the relative brevity of the intervention.

While the main finding shows little real difference in gains between the treatment and control groups of schools overall, the more nuanced analyses provide important insights into the possible conditions under which meaningful change, what Hopkins (2003) described as "improvement for real", could occur. The analysis suggests that the more extensively teachers participated in the intervention (as measured by the number of training sessions attended), and the higher their commitment or enthusiasm (as measured by the number of lessons covered and assessments administered), the stronger was the programme's effect on their learners' spelling and language performance. An added insight that emerges, one that will require new studies to confirm, is the differential impact of individual coaches. The RCUP findings suggest that while instructional infrastructure (Cohen, 2011), in the form of lesson plans, learner resources, and coaches, may be necessary conditions for improvement, the quality and effectiveness of individual coaches may be an often hidden but powerful factor.

Besides this strong finding, the study has also provided substantial evidence around a range of themes. These include further evidence of the serious underperformance in English as a First Additional Language at the start of the Intermediate Phase, and the scale of the performance gap between the genders.

The study pretest dataset suggested that the Grade 4 learners' English-language proficiency is very weak. Pinetown was selected as one of the higher-performing districts in the province, as indicated by the ANA scores. Our findings, however, suggest that there is a significant discrepancy between performance levels, as indicated by the ANA scores, and proficiency levels, as measured by our test. The divergent performance measures may be a function of the different test instruments, or of the different conditions under which the tests were administered and marked.

Another major insight from the pretest analysis is the large performance gap between boys and girls. This gap is evident in both the pretest and the posttest, and is consistent across the study tests and

the ANA results. This trend, identified by Perry (2006) in the early 2000s, and recently confirmed by Zuze (2014), is not adequately understood.

5. Conclusion

The Reading Catch-Up Programme has been shown to have little educationally significant impact. The results of this study, robust as they are, do not suggest any specific policy or programme warrants. The lesson, however, for policy makers is that policy or programme effectiveness claims can and should be tested using robust counterfactual studies prior to system-wide rollout.

The study demonstrates the value of counterfactual research. If this study had used a simple pretest-posttest design (as was used in the initial study), the conclusion would be a *false positive*, namely that the intervention was highly effective. Having a randomly selected control group to provide a valid estimate of the counterfactual allowed us to observe similar gains for the control group, and, by extension, the conclusion is that improved performance cannot simply be ascribed to the intervention. The study also shows the value of conducting replication studies, to address questions of external validity. Assuming that the results of the Gauteng study were reliable and valid, this study demonstrates that policy transfer cannot automatically be assumed. This may be because of substantive differences in language context and language practice across provinces.

Finally, while the study was explicitly designed as an impact evaluation, the data collected for the study are likely to be the basis for a number of additional secondary studies. Thanks to a generous grant from the National Research Foundation, a number of graduate students are likely to undertake more fine-grained analyses of English First Additional Language literacy acquisition in our schools.

References

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics. An empiricist's companion*. Princeton, NJ: Princeton University Press.

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

Educational evaluation and policy analysis. (2007) 29/1.

Fleisch, B. (2008). *Primary education in crisis: Why South African schoolchildren underachieve in reading and mathematics*. Cape Town: Juta.

Fleisch, B., Taylor, N., Herholdt, R., & Sapire, I. (2011). Evaluation of Back to Basics mathematics workbooks: A randomised control trial of the Primary Mathematics Research Project 1. *South African Journal of Education*, *31*(4), 488-504.

Fleisch, B. (2013a). Change at the instructional core: Insights from the Intersen English Catch-Up Programme. Paper presented at the South African Education Research Association meeting, January 2013.

Fleisch, B. (2013b). System reform and primary literacy: Implication for school leadership. In I. R. Haslam, M. S. Khine, & I. M. Saleh (Eds.), *Large scale school reform and social capital building*. New York: Routledge.

Hellman, L. (2012). GPLMS Intersen Catch-up Programme: Analysis of Results. Memo.

Hopkins, D. (2001. *School improvement for real*. London: Routledge/Falmer.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science , 340*, 297-300.

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242-252.

National Education Evaluation and Development Unit.(2013). Report of the National Education Evaluation and Development Unit, Pretoria.

Perry, H., & Fleisch, B. (2006). Gender and educational achievement in South Africa. In V. Reddy (Ed.), *Marking matric: Colloquium proceedings* (pp. 107-126). Cape Town: HSRC Press.

Pretorius, E. J. (2014). Supporting transition or playing catch-up in Grade 4? Implications for standards in education and training. *Perspectives in Education*, *32*(1), 51-76.

Pritchett, L., & Beatty, A. (2012). The negative consequences of overambitious curricula in developing countries. *Center for Global Development Working Paper*, (293).

Raudenbush, S. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*(1), 25-31.

Robinson, D., & Levin J. (1997). Reflections on statistical and substantive significance, with a slice of replication *Educational Researcher, 26*(5), 21-26.

Spaull, N., & Kotze, J. (Forthcoming). Starting behind and staying behind in South Africa: The case of insurmountable learning deficits in mathematics. *International Journal of Educational Development*.

Taylor, N. (2007). Equity, efficiency and development in South African schools. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement*. Dordrecht, Netherlands: Springer.

Taylor, N., Van der Berg, S., & Mabogoane, T. (Eds.). (2013). *Creating effective schools.* Cape Town: Pearson.

Van Staden, S., & Bosker, R. (2014). Factors that affect South African reading literacy achievement: Evidence from prePIRLS 2011. *South African Journal of Education*, *34*(3), 01-09.

Zimmerman, L., & Smit, B. (2014). Profiling classroom reading comprehension development practices from the PIRLS 2006 in South Africa. *South African Journal of Education*, *34*(3), 01-09.

Zuze, T. L., & Reddy, V. (2014). School resources and the gender reading literacy gap in South African schools. *International Journal of Educational Development*, *36*, 100-107.

Annexure

Qualitative information on implementation

As part of the data collection for the Reading Catch-Up Programme randomised control trial, the research team undertook a qualitative data-collection exercise. The purpose of the qualitative data-collection exercise was (a) to monitor the implementation of the intervention, and (b) to gather qualitative information on classroom practices and teacher perspectives. The following questions guided the site visit:

(a) To what extent is the intervention impacting on instructional practice across the range of intervention classrooms?
(b) What contextual factors are impacting on implementation?
(c) How are teachers experiencing the intervention, and what are they saying about it?
(d) What are teachers' attitudes towards the intervention?
(e) What factors may be impacting on the posttest data-collection process?

To address the guiding questions, the research team planned a two-day site visit. The service provider was requested to identify four schools in two of the three district clusters. The research team requested to visit classrooms at the extremes, i.e. very good, and very weak/problematic. The research team also requested that focus group interviews with four teachers be conducted each afternoon during the two-day site visit. Active, articulate, and reflective teachers in the intervention schools were to be selected for the focus group interviews. While the class observations provided insight into the range of implementation levels, the interviews provided insight from best practice. Two members of the research team observed in each of the lessons and participated in the focus group interviews, although one member took the lead with the questioning. The semi-structured questions used in the focus groups centred around four open-ended questions, namely:

- What did you like/dislike about the RCUP?
- How well did the lesson plans work?
- How effective were the various components of the LTSM; and
- What was your experience of the coaching process?

Information about the lessons and the interviews was recorded by hand.

In two of the schools, the site visit team had short interactions with school managers, in one instance with a principal and a head of department, and in the other instance with the head of department only.

The findings section begins with a description of the context of each school and classroom. This is followed by a narrative of the lessons. The second part of this section provides summaries of the experiences and perceptions of the teachers and coaches from the focus group interviews.

Class observations

<u>Primary School X</u>

The school is located in a low-income peri-urban area consisting of self-constructed houses and informal dwellings. The school itself is currently under construction, with new classrooms nearing completion. We had a brief interaction with a head of department, as the principal was not at school, but attending a "meeting". The teacher escorted us to a temporary classroom with 120 learners seated in 11 rows of old desks. According to the teacher, the large number of learners in the class was as a result of a shortage of classroom space, not a shortage of teachers.

The lesson began at 8:36. The "listening and speaking" lesson focused on English-language vocabulary associated with an "At the Hospital" poster, which was displayed on the chalkboard. Part 1 of the lesson began with a recap of the previous "listening and speaking" lesson, using other posters, specifically "At the Beach" and "The Zoo". The Q&A during the recap involved teacher questions and whole-class responses. The teacher's use of code-switching was infrequent. At 8:45 the teacher shifted to the "new knowledge" part of the lesson, namely new vocabulary related to the "At the Hospital" poster. The teacher pointed to one aspect of the poster and asked a question:

> T: Who is reading a message on her cell phone?
>
> L: The girl.
>
> Whole class: The girl is reading the message on her phone.

This process continued for another five or so minutes. At 8:50 the teacher introduced the grammar, or language structure, part of the lesson, which required that learners change sentences from the present simple tense to the future simple tense, using the word "will". The teacher explained the rule for changing present tense to future tense, and then spent the next few minutes going through examples based on the images in the "At the Hospital" poster. The teacher went through the vocabulary words, including "doctor", "nurse", "patient(s)", "cast", and "medicine", and then developed sentences using nouns and verbs. Despite the very large class, the teacher was able to command the attention of almost all the learners and to engage in Q&A. She made limited use of code-switching, mostly restricted to translating individual words, e.g. "*Isiguli* is a patient".

The teacher then repeated a number of examples of how to change present-tense sentences to the future tense, where these examples involved content associated with the activities displayed in the "At the Hospital" poster, e.g. "The doctor talks to the patient > The doctor <u>will</u> talk to the patient". The teacher stressed the need to retain all parts of the sentence and to insert the future tense word "will", but did not deal with changes to the verb itself.

The lesson entered the third part at 9:15, when the children were told to take out their exercise books and complete two sentences that had been written on the board, i.e.

> The nurse will _____.
> The doctor will _____.

The learners struggled to get out their exercise books, as they were sitting on them, as there was no space on the floor for their bags.

<u>Primary School Y</u>

Primary School Y is located within a few hundred metres of a very upmarket suburb, in a peri-urban township area high on a hill overlooking the Tugela River. Although the original school building was built long ago, in 1960, the school infrastructure was in good condition, and the grounds were clean and neat. We were greeted by the principal and the head of department, who had prepared tea and muffins for us. The principal knew a great deal about the intervention and how it was impacting on her teacher's instructional practice. She was particularly proud of her teacher, as she had completed a Master's degree at the University of KwaZulu-Natal, and was considering registering for a PhD. The teacher was one of her former learners at the school, who had chosen to return to her original primary school to teach. The school had classes from Grade 1 to Grade 4, and was what the principal referred to as a "junior primary school".

With a class of 54 learners, the lesson began after the lunch break, at 11:20. The lesson began with a Q&A, with the teacher attempting to elicit learners' knowledge of the English concept of "the future", in preparation for the "listening and speaking" lesson on the future simple tense. The learners were eager to participate, but clearly were not familiar with the English word "future". As the teacher transitioned to the second part of the lesson, she reminded the learners of the language structure/grammar rule that sentences need a "person or people", a "verb", and a "thing".

The main part of the lesson focused on tenses and new vocabulary related to the "At the Hospital" poster. The teacher repeated a range of verbs from the poster, e.g. "carries", "sweeps", "examines", "pays", "X-rays", and "pushes", but in the present tense, and then in the future tense. She introduced a wide range of new nouns to the children from the poster, including "paramedic" and "ambulance". There was a brief Q&A concerning the paramedic and their role in bringing patients to the hospital. This part of the lesson concluded with an informal oral assessment, where the teacher asked the learners to rephrase the sentence "The nurse helps the granny to stand up from the wheelchair" in the future tense.

At 11:45 the lesson shifted into the final part, with the children being required to write down the grammar rule in their exercise books and compose five sentences of their own in the future tense, based on the "At the Hospital" poster. Towards the end of the period, the learners approached the teacher individually to have her mark their sentences. The observers checked the books and noted general mastery of the new knowledge, as evidenced by the high number of grammatically correct sentences produced.

Primary School Z

The school is located in a lower middle-income suburb of Pinetown. The school was built high on a hill overlooking a valley, and is in good condition, with well-maintained grounds. The principal was attending a "meeting", and was not present to greet us. We were met by the teacher who had prepared for our classroom visit. After the lesson, we had a brief conversation with the head of department.

With 62 learners in the class, the lesson began at 8:30 with "listening and speaking" Q&A related to the vocabulary used in a particular story. Holding up vocabulary flashcards, the teacher asked questions such as "What is a needle?" The teacher then focused on a range of English words key to the story, including "tears", "needle", "eyes", "voice", "*gogo*", "proud", "garlic", "better", "wedding", "bridegroom", "dress", and "pins".

At 8:52 the second part of the lesson began, with the teacher reading the story "My Sister's Wedding" (New Heights) aloud to the learners. The learners shared copies of the book. The teacher read several pages in a highly animated manner, and then stopped to ask questions about the meaning of the text, and broader questions about how the learners would feel if they were in the situation related in the text. The learners were engrossed in the story and the teacher's reading. On completing the story, the learners read the story aloud as a class. At 9:05 the teacher began to go through the story to help the learners identify the proper nouns and the common nouns in the story. When they responded to questions about the story, the teacher gently reminded the learners to use the correct gender, i.e. "him" or "her", or "he" or "she".

The final part of the lesson involved the teacher asking the children to engage in an interpretive exercise related to the question of whether or not the learners believed the "facts" in the story to be true, namely that a pinprick can improve a singing voice. This question was very difficult for some of the learners to answer, as they had not encountered a question that required such critical thinking before, but a small group of learners were able to understand that the story was a humorous story rather than a factual story.

Primary School AA

The school is located in a rural section of the Pinetown district. A head of department greeted us and apologised that the principal was attending a "meeting". It was a small school, with only one class per grade, with the class that we observed having only 26 learners. The school had both old and recently renovated permanent structures and temporary classrooms, as well as a new Grade R wing.

The phonics lesson began at 11:02. The teacher began by explaining the principle of the long and the short /a/ sound, and how adding an e to the end of a word changes the /a/ sound from the short to the long sound. She used the example of "tap" and "tape". The learners did not really seem to understand what she was explaining, but sat silently.

Once she had completed her explanation of the principle of the short and the long /a/, she gave a number of examples, including "spade", "make", "late", "shape", and "wave", and then added other words, such as "again". These words were displayed on flashcards. She asked the learners to come to the board to point out various objects in the "At the Beach" poster, including the "spade", and the umbrella, which gave "shade".

At 11:13 the learners were instructed to take out their exercise books and copy down notes that the teacher had prepared before the lesson. The teacher asked the learners to complete two sentences:

1. The children use the _____ to dig.
2. The umbrella make[s] some _____.

The children were slow in copying down the notes, and struggled to complete the two sentences. The lesson ended at 11:42. The teacher commented to the observers that this was a 30-minute lesson.

Teacher focus group interviews and coach interviews

The general sentiment regarding the RCUP was that the programme was very beneficial. The teachers specifically mentioned that they really appreciated the phonics elements, as most of them had not been trained as junior primary school teachers, and did not know about phonics, and had never used it in their teaching. The teachers commented positively on the books. Most of them had never used story books, particularly books at the appropriate reading levels. They liked the ways in which the programme taught comprehension and linked new sets of comprehension questions, the "where", the "what", the "who", and the "why", with the pictures and the texts. They found the flashcards to be very useful, and they really liked the weekly spelling test activity, which they found stimulated friendly competition, even though some children stayed away from school to avoid the assessment. They liked the lesson plans, as they described how these plans ensure that their lessons do not lose focus.

One of the major themes that the teachers mentioned was planning. They really liked the fact that the lesson plans were all prepared for them. They really disliked all the "paperwork" required by their HODs, associated with lesson planning for the normal curriculum.

The main concern or complaint related to the use of time. The first focus group mentioned the difficulty of finishing the lesson within the one hour allocated, and they reported that some of the teachers used the time allocated to other subjects, or even extended the school day. They also mentioned that these lessons took precedence, and that other activities, such as cultural days, had to be postponed to ensure that the lessons were covered. While teachers in the second focus group made a similar observation, one of the teachers in this group described how over time teachers became better at completing the lesson plan tasks within the allocated one hour, as they became better prepared.

One of the teachers described how his teaching had changed from all "theory" to a balance of "theory" and "practice". By theory he meant teachers standing up and imparting content to the learners, and then getting learners to repeat it. By practice he meant that learners got to see pictures, they tried to form sentences themselves, and they read for themselves. Other teachers described the method as making things practical, with more or less a similar emphasis on children being more active in the learning process.

One teacher identified the RCUP as the best programme she has experienced in her entire career. She described herself as being initially very disillusioned by the many changes associated with the curriculum, and very reluctant with regard to the RCUP, but explained how she had changed her mind after she had used the materials and worked with the coach.

Lesson plans

Teachers were very positive about the lesson plans. They described them as "spot on, spot on". They believed that the lesson plans were "very useful" and "very interesting", that the "pacing was okay", and that the lessons were pitched at the "right level". They liked the fact that the lesson plan structure was the same every week. Teachers singled out the reintroduction of the spelling test as a very positive development in the lesson plans.

LTSM

The teachers were equally positive about the LTSM, including the books, the posters, and the flashcards. The teachers spoke positively about specific titles, including *The Little Red Hen*, *Oops*, *Lazy Mandla*, and *Fruit Salad*. The strengths of these books were their excellent illustrations, their use of colour, and their fun, engaging stories. They liked the fact that the stories came with the flashcard words. Both high frequency and the phonics words linked the vocabulary of the stories. The teachers also spoke very positively about the posters, and the fact that the learners were given the posters to colour in. One teacher identified the summaries of the books as an important asset, as they introduced learners to the skill of summarising that will be required in the Intermediate Phase curriculum.

Coaching

The teachers were very happy to have had a coach. They found that the coach was very motivating and helped build their confidence. One of the teachers responded, "This is the first time in 22 years I had someone I could rely on." One teacher described being in a panic the first time the coach came, but settled down almost immediately. The teachers described how the coach helped them prepare for the lessons. The coach assisted with classroom activities, particularly with regard to time management. For example, the coach helped the teachers understand the value of writing the content on a large piece of paper and press-sticking the paper to the board, rather than writing the content on the board during class time. Teachers specifically identified how the coaches assisted with phonics and the pronunciation of sounds, which was something that none of them had ever done before.

The coaching process made teachers feel more confident, because they were criticised constructively, and because the coaches knew more about the teaching process than did the teachers. Teachers raised concerns about the competency of many of the heads of department, and the general sentiment was that they did not find the visits of the district facilitators very helpful. What was surprising was that none of the teachers raised concerns about being observed while teaching, or receiving critical feedback. Instead, they found both to be very helpful when it was done by the coaches.

Resistant teachers

One coach identified two out of the 20 teachers as actively resisting the programme. The other coach identified only one of the teachers as actively hostile to the programme. The gender and age profile of the resisters is not consistent with the pattern found in Gauteng.

On the basis of the classroom observations and the interviews, we have a number of insights. Given that many of these insights are drawn from the focus group interviews, they tend to be biased towards "good" practice, rather than "typical" practice.

- It was evident that the teachers had prepared for the observations, and the lessons demonstrated that the teachers understood how to translate the lesson plans into classroom instruction. The pattern of exercises completed in the learners' exercise books provides evidence that the teachers had been following the lessons more or less as required. There were, however, some discrepancies, where at least one teacher was two weeks behind the scheduled lessons.

- The quality of the actual lessons observed differed considerably. Although all observed lessons followed the lesson plans, and made use of the LTSM provided, the pacing and sequencing of the parts of the lesson differed. Two lessons were less than satisfactory, the poor quality of one of the lessons possibly being attributable to the very large size of the class (more than 100 learners).
- One lesson could be described as exemplary. The teacher followed the lesson plan faithfully and animated the classroom instruction in such a way as to emotionally and intellectually engage the learners in the class. The teacher ended the lesson with a higher-order thinking activity, which really challenged the learners.
- Overall, the lesson observations showed that the lesson plans and the LTSM were appropriate for both the teachers and the learners in the province.
- The general sentiment of the interviewed teachers was positive with regard to the programme as a whole, and the various components thereof, including the lesson plans, the LTSM, and the coaching.
- The lesson plans significantly extended the instructional repertoire, adding knowledge and skills particularly in the use of phonics for decoding, as well as the weekly spelling test.
- The teachers found the LTSM very effective. They singled out the posters and related questioning techniques as very useful for the "listening and speaking" component of the First Additional Language curriculum. They found that the flashcard words linked to the poster provided the learners with multiple opportunities to understand the meaning and spelling of high-frequency words.
- The graded readers were also identified as a very positive set of resources. The teachers found that the learners engaged with the pictures and the interesting story narratives. The learners felt empowered by having reading books at their particular level of fluency.
- The teachers were very positive about the coaching process, and they indicated that it was critical at the start of the change process to "empower" teachers to begin using the new programme. The coaching also provided the teachers with key insights into specific aspects of the programme, particularly with regard to the pronunciation of words and phonics. The teachers did, however, recognise that such coaching may not be necessary in the long term, particularly if the heads of department and the district facilitators have a good understanding of the programme and can add value.

Endnotes

[1] We use the terms "reading" and "literacy" interchangeably. While the RCUP programme was clearly geared to improving reading proficiency, the test instrument was oriented towards the measurement of certain literacy skills, rather than oral reading fluency and comprehension.

[2] The terminology of "treatment group" and "control group" originates from the literature on medical trials, where a particular drug, or "treatment", was being trialled. The terminology is now widely used across fields in impact evaluations. We use the terms "intervention group" and "treatment" group interchangeably.

[3] A paper by Kremer, Brannen and Glennerster (2013) provides a concise review of international RCT studies focusing on education.

[4] Initially, we tried to select schools based on the original below ANA 50% level and between 30 and 90 learners criteria. But in order to find 100 schools, we had to start relaxing some of these criteria. Read the full sampling report in the pre-analysis plan, to see exactly what we did.

[5] The power of the statistical test refers to the probability of avoiding a Type II error (that is, incorrectly rejecting a null hypothesis). Therefore it represents the likelihood of drawing the correct conclusions about the significance of differences between groups. Typically, a power level of 80% is considered high enough to detect differences, while keeping sample sizes reasonable.

[6] The ICC is the proportion of the total variation in test scores that is accounted for by between-school variation; the remainder is accounted for by within-school variation among learners. It describes the level of inequality between schools. The higher the ICC, the larger are the systematic differences in achievement scores between schools, and the more groups are required in the sample.

[7] In order to determine appropriate sample size, it is necessary to have some prior knowledge of expected size of the intervention effect. In much of the contemporary US-based literature this has been standardised to a common effect size unit, that is, percentage of the standard deviation of the outcome measure. This allows for comparison across studies using different scales. While the original PRMP study did not report results in percentage of the standard deviation of the outcome measures, the percentage point gains reported were very high. The use of 0.2 standard deviations can be regarded as a moderate effect size relative to those typically observed in the international literature on school interventions.

[8] Given this core finding, the question of cost-effectiveness is of no consequence.

[9] For an explanation of instrumental variables and two-stage least squares regression, see Angrist and Pishke (2009).

[10] An internal Department of Basic Education analysis of the Annual Survey of Schools indicated this.

[11] Although no baseline score is inserted as a control variable, there is no reason to expect substantial baseline differences between treatment and control schools because of randomisation. Furthermore, we include the stratification dummies in the regression, to further control for differences in school characteristics, including prior ANA achievement, which was one of the dimensions influencing stratification.

[12] Although Pretorius' study used only one school, and therefore should not be considered as a benchmark for typical effect sizes.

[13] Lemons, Fuchs, Gilbert and Fuchs (2014) describe similar patterns with strong gains in counterfactual study groups. Their account, however, stresses the shifts in the entire school system as a result of improved early reading teaching.

APPENDIX D: CHAPTER 4 RCUP Pre-analysis Plan

READING CATCH-UP PROJECT: THE IMPACT OF AN 11-WEEK TEACHER SUPPORT PROGRAMME ON PUPIL READING PROFICIENCY

Pre-Analysis Plan

Brahm Fleisch, Thabo Mabogoane, Volker Schöer, Stephen Taylor

16th June, 2014

A. Introduction to the study

We are undertaking a randomised controlled trial of a Reading Catch-Up Programme amongst fourth grade pupils in South Africa. The study will evaluate the efficacy and cost-effectiveness of an eleven week programme which focuses on improving the performance of fourth grade students in English, which is the second language for the majority of targeted students. The study will be conducted in the Pinetown district of KwaZulu-Natal Province in South Africa.

As is the case for the majority of students in South Africa, those in the targeted schools experience a transition in the language of instruction at the fourth grade. In grades 1 through 3 most schools teach in the Home Language of students (isiZulu in the case of Pinetown) and then transition to English as language of instruction in grade 4. Research has demonstrated that most South African children in poor communities have accumulated large learning deficits in English proficiency by the fourth grade (Taylor, 2011).

The intervention consists of three components: scripted lesson plans, additional reading resources and in-class instructional coaching and training. The study randomly assigned 40 schools to the treatment group and 60 schools to the control group. Baseline testing was conducted between the 8th of April and the 18th of April 2014. Follow-up testing will be administered between the 17th of June and the 27th of June 2014. This study has the potential to contribute significant policy insights with regard to the value and efficacy of large-scale remediation strategies in the public school sector.

Section B provides further motivation for the study and describes the content of the intervention. Section C describes the sample of students and schools we use for the study. Section D describes the key data sources to be used in analysis. Section E outlines our various hypotheses that we plan on testing. Section F describes our estimation strategy. Section G discusses various robustness checks we plan on conducing in order to assess the sensitivity of our results to issues such as attrition. Section H outlines a cost-effectiveness comparison we will conduct.

B. Motivation of study and content of the intervention

While the body of theory on system-wide change has gained considerable influence, there is limited supporting robust empirical evidence. Although the number of studies that estimate causal links between interventions and system-wide changes in learner performance has grown since Glewwe's (2002) well known review, the body of research is modest (McEwan, 2013). Lucas et al (2013) in their literature review cited five published studies in the past ten years. These include a study of remedial tutoring for low-achieving learners in India (Banerjee et al. 2007); an Indian study of scripted lessons (He et al, 2008); and work on structured lesson plans and in-service training (Friedman et al (2010), Piper & Medina, 2011). Most recently, Piper et al (2014) have published positive findings of an RCT study that made use of scripted lessons, high quality materials and in-classroom support in English and Kiswahili.

Theory of change

1. While a range of factors influence learning outcomes, including school and non-school variables, instruction or instructional practice is one major influence on learning.

2. Instructional practice is the process of teachers making use of learning resources in a structured way with learners around particular content.

3. One of the key characteristics of South African education is that the dualistic nature of learning outcomes is mirrored by dual types of instructional practice (e.g. Hoadley, 2010).

4. It is likely that weak instructional practices have a causal impact on learning outcomes in the poorly performing part of the school system.

5. To substantially shift achievement in the weak part of the schooling system it may be necessary to apply a comprehensive instructional change intervention, involving a set of coherent and aligned instructional inputs.

6. For this intervention, the instructional inputs include scripted lesson plans, aligned learning materials and in-classroom support to teachers.

7. The scripted lesson plans provide specification of the new instructional practice including faster paced instruction, more appropriately sequenced content, and dramatically expanded pedagogic repertoires.

9. Specifically in primary school teaching reading in the First Additional Language, the new expanded repertoires include systematic teaching of phonemic awareness and phonics; strategies that focus on increased reading speeds or fluency; guided reading strategies; vocabulary development and strategies that improve comprehension.

10. The role of the learning materials is to provide the appropriate resources to ensure that learners are able to develop and consolidate knowledge and skills related to reading fluency, vocabulary development and guided reading. Twelve sets of graded reading books are provided.

11. The role of in-class support is to fuse capacity-building and accountability. The in-class support allows for modelling of the new practice on site and the gradual development of teachers in the new practice from novice to expert. The in-class support also allows teachers to manage the emotional labour, i.e. stress, insecurity and anxiety associated with developing a new professional practice mid-career. Finally, the in-class support allows for the development of professional accountability (rather than bureaucratic or performance accountability) in an environment of trust. There are 5 scheduled visits to each school from a specialist reading coach over the 11-week period.

12. This is the general theory of change, and applies to the Reading Catch Up Programme (RCUP), though it has an additional component – a strong remedial focus. It assumes that many learners in Grade 4 have not had curriculum level teaching in the first three grades, particularly in FAL English. A catch-up programme assumes that because of the physiologic/chronologic development and the potential transfer of some literacy skills from the Home Language to the First Additional Language, it is possible to 'fast-track' and consolidate reading practice over a relatively short period.

13. A common educational problem in South Africa and in many developing countries is that the levels of learning lag behind the concepts required in the curriculum (e.g. Banerjee and Duflo, 2011). Given that all learning builds on prior learning this creates an imperative for remedial programmes that aim to consolidate concepts and skills that should have been mastered during earlier phases of schooling. Remedial reading programmes such as that described in Banerjee *et al* (2010) have shown success in RCT's elsewhere in the world.

14. While the Catch-Up Programme cannot address all the learning challenges, particularly for learners with organic and severe learning disability, it promises to strengthen English reading performance for learners in the middle to lower range of performance.

C. Description of the sample to be used in the study

Deriving a Sampling Frame

We used information from the Department of Basic Education's Annual National Assessments (ANA) datasets of 2012 and 2013 (grade 4 only) to obtain a sampling frame of 100 schools. We also identified 4 replacement schools, which we had planned to draw on if schools were unavailable.

After generating the necessary information at the school level (e.g. enrolment in each year, mean language performance, etc) and merging the 2012 and 2013 ANA datasets into one, we deleted all schools except Public schools in the Pinetown district. This left 329 schools.

We then selected schools into the sampling frame in 3 steps:

1) The first batch of **46 schools** to make it into the sampling frame are ones we are fairly sure about from both 2012 and 2013 data. It includes schools that had enrolment of at least 15 grade 4 students in both 2012 and 2013, had an average score of less than 55% in 2013 language, had an average score of less than 50% in 2012 language, had at least 5 learners writing English as a First Additional Language (FAL) in 2013, had at least 50% black learners,

had no more than 115 grade 4 learners in 2012 and no more than 115 grade 4 learners in 2013, and are not quintile 5 (the most affluent of 5 official socio-economic categories of South African schools).

2) The next batch of **36 schools** enters as we relax the condition that schools must have 2013 FAL marks (i.e. we can include schools that wrote English on the Home Language level), but we tighten the proportion of students that needs to be black to 70% and lower the performance cut-off to 50% in both years: So this included all schools that had enrolment of at least 15 grade 4 students in both 2012 and 2013, had an average score of less than 50% in 2013 language, had an average score of less than 50% in 2012 language, had at least 70% black learners, had no more than 115 grade 4 learners in 2012 and no more than 115 grade 4 learners in 2013, and are not quintile 5.
3) The next batch of **18 schools** get in through ignoring 2013 requirements (though still insisting that there is data for both 2012 and 2013), but we now insist on schools taking FAL and we raise the upper enrolment limit from 115 to 120: So we now include all schools that had an average score of less than 50% in 2012 FAL, had at least 15 learners in 2012, had no more than 120 learners in 2012, and are not quintile 5.

This yielded 105 schools, but 1 school had an average score in 2013 of zero so we chose to delete this suspicious case. This lead to a sampling frame of 104 schools and the plan was to identify 40 treatment schools, 60 control schools and 4 replacement schools. We arrived at the exact parameters of the 3 batches in an iterative process aiming at getting to a sampling frame of 104 schools – the most appropriate 104 schools in the Pinetown district to receive this intervention. In other words, I set the various restrictions as strictly as possible so as to have 104 schools. Initially, we had intended on using only schools with less than 40% average. But we had to relax this somewhat to get as many as 104 schools, given that we excluded other schools on the basis of having too few or too many learners, or belong quintile 5, etc.

By design therefore, the schools in our sampling frame are relatively poor (though not as poor as some deep rural South African schools), majority not home language English, neither especially small nor especially large, and relatively low-performing. This ensures the maximum possible external validity for the research pertaining to the large under-performing section of the schooling system at the top of the policy agenda, subject to the constraint that we are working in a single education district.

Stratification of the sampling frame and randomisation

The first step after identifying the 104 most suitable schools for inclusion in the project was to select 4 schools using a computerised lottery to be the replacement schools. After taking them out of the sample, we were left with the 100 schools.

With randomisation there is no *a priori* reason to expect a lack of balance between treatment and control groups. However, stratifying the sampling frame can ensure that any differences on baseline characteristics are minimal and can have the added benefit of leading to narrower confidence

intervals when conducting the eventual analysis. In order to ensure a good balance between treatment and control schools the following process of stratification was followed:

1) Two equal-sized categories of school socio-economic status (SES) were identified. The higher SES category included all 39 quintile 4 schools as well as 11 randomly selected quintile 3 schools. The lower SES group of 50 schools therefore included the remaining 46 quintile 3 schools and 4 quintile 2 schools.
2) Within each of the two SES groups, two equal-sized groups were identified to distinguish between smaller schools and larger schools. Consequently, we have 4 groups of 25 schools: high SES larger schools, high SES smaller schools, low SES larger schools and low SES smaller schools.
3) Within each of the 4 groups, schools were then ranked on their language performance in ANA 2013. This means that each school has a rank position of between 1 and 25. Within each of the 4 groups, schools were then split into groups of 5 on the basis of their rank. This results in 20 separate groups or "strata". For example, strata number 1 includes the 5 worst-performing large high SES schools. Strata number 20 at the other extreme includes the 5 best-performing small low SES schools.

Using a computerised lottery we then selected 2 treatment schools and 3 control schools from within each of the 20 strata. This produces 40 treatment schools and 60 control schools, which should by construction be closely balanced on school size, quintile and baseline performance.

We then wanted to make sure that the sample was more or less balanced across the 4 education circuits of Pinetown. We set a rule that no circuit could have less than 30% treatment schools or more than 50% treatment schools. If this rule was violated the plan was to re-randomise by increasing the seed number by 1000. The second time round we got to a sample that met this condition.

Unfortunately, just prior to baseline testing, the district officials requested to remove 3 control schools from the project. The reasons provided were legitimate and would have applied equally to treatment schools had it been necessary, e.g. a school was not in fact part of the Pinetown district (this would probably have been caused by an error in the administrative data we used to derive the sampling frame). The district office immediately replaced these 3 schools with 3 new control schools. It was too late to re-randomize or to recommend using any of our replacement schools. Fortunately, we have baseline data on these schools, so even if these new control schools are systematically better or worse performing than the original 3 schools we lost, there is no strong reason to expect them to be on a different change trajectory. Initial analysis indicates that the 3 control schools that we lost were slightly better performing than average (looking at ANA results); similarly the 3 control schools that we gained were also better performing than average (looking at our baseline results). So at least we did not systematically remove weak control and replace them with strong control schools (or vice versa), which would have caused a potential bias. Nevertheless, we have a plan to test the robustness of our results to possible bias introduced by this interference with the randomisation, which is outlined in Section G.

D. Key data sources

The study uses a number of data sources. Predominantly, the study uses the baseline and endline literacy test to evaluate the impact of the intervention. The baseline test occurred at the beginning of the second academic quarter (between the 8th of April and the 18th of April 2014) while the endline test will be administered in the last week of the quarter (between the 17th of June and the 27th of June 2014). The intervention programme occurs throughout the second quarter.

An independent service provider is contracted to conduct the baseline and endline evaluations, and to undertake the data capturing. A different organisation is implementing the intervention programme. The service provider undertaking the field testing and data capturing was blind, i.e. they were not informed if the schools were part of the intervention group or part of the control group.

Baseline and Endline tests:

The baseline literacy test contains 35 items, spread across the following domains: spelling (20 items), comprehension (6 items), language (6 items), and writing (3 items). The endline literacy test is exactly the same as the baseline test, except for four additional items (2 spelling items and 2 picture comprehension items). These are particularly easy items and were added due to the fact that baseline scores showed a noticeable floor effect, with high proportions of students achieving very low scores. The inclusion of these easy items should enable us to identify greater variation of achievement at the low end of the distribution, thus improving our chances of picking up a treatment effect (if indeed the intervention influences achievement at the lower end of the performance distribution).

Baseline and Endline teacher questionnaires:

On the same day as the student testing, the fieldwork agency administers questionnaires to teachers. This occurs at baseline and endline. These collect information on teacher characteristics, such as demographic information and information on qualifications and experience. They also collect information, albeit limited information, on aspects of teacher belief and behaviour that we expect might change in response to the intervention.

Baseline and Endline student characteristics:

At the start of the literacy test, students are required to complete information on 4 things: gender, age, exposure to English at home, and extent of reading activities in the home.

Administrative data:

We will use administrative data from the Annual Survey of Schools (ASS) on school characteristics and administrative data on school performance in literacy and numeracy (ANA 2012, 2013 and 2014). The ANA of 2014, which is to be administered in September 2014, will provide an alternative outcome measure.

Reports by coaches and implementation agent:

We will receive some information on implementation (e.g. number of visits from coaches and extent of curriculum coverage) from the Implementing Agent.

Qualitative data from site visits and focus groups:

To monitor implementation and to ascertain the range of teachers' perspectives about the intervention programme, we conducted a two day site visit. This visit included four lesson observations in four schools selected at either end of the implementation spectrum: full fidelity - no implementation. We also conducted two focus group interviews with implementing teachers identified by the service provider. Teachers were selected on the basis of their level of articulateness about the programme.

E. Hypotheses to be tested

Using the information from the endline evaluation (literacy test), the learner and teacher questionnaires, and coaches' reports, we will test a number of hypotheses regarding the impact of the instructional change intervention on learner performance and on teacher approaches to teaching reading.

Impact on learning outcomes:

Hypothesis 1: *The RCUP has a positive average impact on Grade 4 learners' endline evaluation scores (literacy test).* The likely transmission mechanism of this improvement in the endline achievement of learners in treated schools relative to learners in control schools is based on better teaching by teachers who were exposed to the instructional change intervention of the RCUP study.

We will test this hypothesis for the learners' overall test scores as well as for the five specific literacy domains (*Spelling, Comprehension, Language, writing and Picture Comprehension*) contained in the literacy test. As the RCUP intervention might impact different literacy competencies, performing the analysis for the five literacy domains will allow us to unpack the possible impact on nuanced learning outcomes.

The impact on outcomes will be tested using the following individual indicators from the baseline and endline literacy test:

Overall score:

- Overall score calculated as the percentage score of correctly answered questions in the literacy test (baseline and endline test, all questions from sections A, B, C, D and E)

Five literacy domains:

- Percentage score calculated as the percentage score of correctly answered questions in Section A (*Spelling*) of the literacy test (baseline and endline test, all questions from *Section A: Spelling*)
- Percentage score calculated as the percentage score of correctly answered questions in Section B (*Comprehension*) of the literacy test (baseline and endline test, all questions from *Section B: Comprehension*)
- Percentage score calculated as the percentage score of correctly answered questions in Section C (*Language*) of the literacy test (baseline and endline test, all questions from *Section C: Language*)
- Percentage score calculated as the percentage score of correctly answered questions in Section D (*Writing*) of the literacy test (baseline and endline test, all questions from *Section D: Writing*)
- Score out of 2 in Section E (*Picture Comprehension*) of the literacy test (only included in the endline test, all questions from *Section E: Picture Comprehension*)

Heterogeneous treatments:

Hypothesis 2*: Full coverage of the RCUP yields the largest impact on learning outcomes of Grade 4 learners.* The coaches record information on the number of tests administered by teachers and the coverage of the RCUP curriculum, on a week-by-week basis. Therefore, within treatment schools there should be some variation in the extent to which lessons were delivered and content was covered by teachers. We also have information on the number of visits from coaches (though it is not expected that this should vary substantially due to contractual agreements with the service provider). We can derive a new treatment variable, which is 0 for control schools and continuous (or categorical) amongst treatment schools.

Hypothesis 3: *Full coverage of the RCUP yields the largest impact on teachers internalizing key pedagogies of teaching how to read.* Teachers are more likely to internalize key pedagogies if more fully exposed to the RCUP programme. The data sources and variable construction will be similar to that outlined in Hypothesis 4. The same set of outcomes as discussed under Hypothesis 2 will be used here.

Hypothesis 4: *The effect of treatment is different depending on the individual doing the coaching.* The RCUP is administered in 40 schools and these schools are split between 2 specialist reading coaches. One can therefore regard the treatment group as consisting of two separate treatment arms – each with a different coach. Although we do not have sufficient statistical power to be confident of precisely measuring whether the treatment effect is different depending on the coach it is certainly something we will investigate.

Impact on intermediate outcomes

Hypothesis 5: *The RCUP has a positive average impact on Grade 4 teachers' attitudes and approaches to teaching literacy/reading.* By testing whether treatment assignment is linked to changes in various intermediate outcomes such as teacher attitudes and approaches, we aim to understand the mechanisms through which the RCUP has an impact (or not). A key part of the RCUP programme is to change teachers' way of instruction by emphasising key pedagogical approaches to reading. These approaches form part of the intervention and are repeatedly emphasized by the coaches and in the lesson plans. In particular, a key objective of the RCUP is to get teachers to internalize these core pedagogies to affect long term instructional changes in teaching.

We will investigate the following sub-hypotheses:

- *Did the intervention change how teachers ranked the Importance of different types of learner activities when teaching reading?* (We use the following question from the teacher questionnaire: "How important do you consider the following learner activities to be in the teaching of reading?" answers: 1) working in pairs; 2) Independent reading in class; 3) preparing projects or posters to be shown to the class; 4) taking reading materials home to read; 5) homework assignments; and 6) Tests, examinations, other assessments"). Similarly, we ask the following question: "In your opinion, which of the following are the most important three priorities when teaching reading?" answers: 1) Critical Thinking; 2) Spelling; 3) Whole class reading aloud; 4) Phonics and letter blends; 5) Whole language; and 6) Comprehension"). The RCUP focuses on Spelling, Phonics and letter blends, and Comprehension. We expect that teachers who got exposed to the RCUP should have internalised the importance of these three areas and identify them as key priorities in teaching reading. This data was not collected in the baseline teacher questionnaire, but given randomization any significant differences on the endline can be attributed to program impact.
- *Did the intervention change the determinants of teacher job satisfaction?* The role of the coach is in part to play an inspirational role. This could mean that teachers are inspired to care more about learner progress than about other conditions of service. We ask teachers at baseline and endline how important they believe each of the following 5 factors are in determining their job satisfaction: a) quality of school buildings, b) salary level, c) seeing my learners learn, d) availability of classroom resources, e) opportunities for promotion.
- *Did the intervention change the frequency with which learners were able to take books home?* One of the components of the RCUP is the additional graded reading booklets. This may have led to learners more frequently having reading materials to take home. In both baseline and endline, we ask teachers how frequently learners take books home with them.
- *Did the intervention affect the time spent by teachers on lesson preparation?* Since one of the components of the intervention is to provide teachers with clearly scripted lesson plans, we may actually observe a negative effect on lesson preparation (at least in the short run). Nevertheless, it will be interesting to see if teacher responses to questions about lesson preparation changed noticeably between treatment and control schools.
- *Did the intervention change teacher attitudes towards school improvement interventions?* We ask teachers the extent to which they find school improvement interventions to be helpful.

- *Did the intervention change the frequency with which teachers administer tests?* Regular assessment of children is one potentially important component in teaching reading, and it is a focus of the RCUP. We ask teachers in baseline and endline surveys how often they administer written language tests.
- *Did the intervention change teachers beliefs/aspirations about when a child can reasonably be expected to achieve reading fluency in a) their home language and b) English?* International assessments such as PIRLS 2006 indicate that the majority of South African children had not learned to read for meaning by grade 5. Anecdotal evidence suggests that many teachers do not believe it is possible for children from poor socio-economic backgrounds to learn to read effectively in the early grades. The inspirational role of the coaches may lead to a greater sense of self-efficacy amongst teachers and greater optimism about the possibility for young children to learn to read in the early grades. In both baseline and endline, we ask teachers, "by what grade do you think it is possible for the learners in your school to reach reading fluency in their home language and English?"

A key limitation for the analysis of the second hypothesis will be the small number of observations. We generally only observe one teacher per school with a total of 40 treated and 60 control schools. Hence, we might not have enough power to establish any statistically discernible differences in these intermediate outcomes. A further limitation is that the majority of intermediate outcomes rely on self-reported data from teachers.

Heterogeneous treatment effects:

Hypothesis 6: *Learners, teachers and schools with varying characteristics are likely to experience differences in the magnitude of impact of the RCUP on learning outcomes.* For instance, male teachers may react differently to the intervention compared to female teachers; better instructions by teachers may impact more depending on learner gender; or the treatment may impact differently depending on baseline achievement. In order to examine treatment heterogeneities, we will investigate a number of learner, teacher and school characteristics obtained through the learner questionnaire (four questions in the baseline literacy test), the teacher questionnaires (baseline and endline) as well as school level characteristics obtained from administrative data sources.

At the individual learner level:

1. Does treatment impact vary depending on baseline achievement (percentage score calculated from the baseline literacy test)?
2. Does treatment impact vary depending on learner gender (demographic question asked on cover sheet of baseline evaluation: options Boy/ Girl)?
3. Does treatment impact vary depending on learner age (demographic question asked on cover sheet of baseline evaluation: age options given from 7 – 13; with additional option "Older")?
4. Does treatment impact vary depending on reading activity in the home (question asked on cover sheet of baseline evaluation: "Do adults in your home read books, newspapers or magazines?" answers: *Never; Sometimes; Often; All the Time)?*
5. Does treatment impact vary depending on whether English is used as a medium of conversation at home (question asked on cover sheet of baseline evaluation: "*Do you speak English at home?"* answers: *Never; Sometimes; Often; All the Time).*

At the teacher level:

1. Does treatment impact vary depending on the size of the class taught by teacher (baseline Teacher questionnaire, Q1)?
2. Does treatment impact vary depending on the gender of the teacher (baseline teacher questionnaire, Q2)?
3. Does treatment impact vary depending on the age of teacher (baseline teacher questionnaire, Q3)?
4. Does treatment impact vary depending on teacher qualifications (baseline teacher questionnaire, Q4)?
5. Does treatment impact vary depending on years of teaching experience (baseline Teacher questionnaire, Q5)?
6. Does treatment impact vary depending on how many times teachers have attended additional in-service courses over the last three years (baseline teacher questionnaire, Q6 for number of courses, and Q7 for total number of days)?
7. Does treatment impact vary depending on baseline availability of reading books (baseline teacher questionnaire, Q8 options 5 & 6 "availability of: 5) Bookshelves; 6) A classroom library, book corner or book box")?
8. Does treatment impact vary depending on baseline ability of learners to take books home (baseline teacher questionnaire, Q10 "Are learners able to take reading books home?")?
9. Does treatment impact vary depending on baseline teacher effort levels in preparing lessons (baseline teacher questionnaire, Q13 "How many hours, on average, do you spend in a typical school week working on lesson preparation for this school?" and Q14 "Do you prepare your own lessons?")?
10. Does treatment impact vary depending on the teacher's own attitude/liking for reading (baseline teacher questionnaire, Q25 "How many books have you read for pleasure so far this year?")
11. Does treatment impact vary depending on the teacher's initial assessment of learners' reading ability (baseline teacher questionnaire, Q29 "In your opinion, what proportion of your Grade 4 class could read fluently in their home language at the start of Grade 4?")? This will be tested over and above the actual level of ability of learners (as measured on the baseline test). It is therefore a test of whether teacher perceptions of learning problems affect the teacher's receptiveness to the intervention.
12. Similar to (11), we ask does treatment impact vary depending on the extent to which teachers report that learners struggle with the transition in language of instruction from Mother Tongue to English.
13. Does treatment impact vary depending on how many other non-governmental school improvement interventions the school has participated this year?
14. Does treatment impact vary depending on the teacher's initial aspirations/beliefs about when a child can reasonably be expected to achieve reading fluency in a) their home language and b) English?
15. Does treatment impact vary depending on the teacher's initial response on what determines their job satisfaction? It may be that teachers who cared less about learning outcomes would have been less receptive to the intervention. Alternatively, it may be that a greater

impact occurs for teachers who initially did not care enough for learning outcomes, but whose priorities changed due to the intervention.

16. Does treatment impact vary depending on the amount of time per week spent by teachers on teaching in the classroom?
17. Does treatment impact vary depending on teacher attitudes to classroom support from a) district officials? And b) other non-governmental interventions?
18. Does treatment impact vary depending on how often the School Principal or Head of Department advises teachers on their teaching?
19. Does treatment impact vary depending on the level of parent involvement in schools? It may be that coaches fill an accountability gap that is left by low parent involvement, thus leading to greater impact.
20. Does treatment impact vary depending on the teacher's initial frequency of administering written tests in class?

At the school level:

1. Does treatment impact vary depending on initial average performance level of the school (school average percentage score in baseline literacy test)?
2. Does treatment impact vary depending on the socio-economic status of the school (only quintiles 2, 3 and 4 schools are in the sample)?
3. Does treatment impact vary for urban versus rural schools?
4. Does treatment impact vary depending on the school size (enrolment in Grade 4 and total number of learners enrolled in school for Grade 1- Grade 6)?
5. Does treatment impact vary depending on the language of instruction in the Foundation Phase (Grades 1 to 3)?

It could be that interactions of some of the above characteristics might exhibit different treatment effects. We will test a number of these interactions including, among others, learner and teacher gender interactions, learner gender and baseline achievement; learner gender and English as medium of conversation at home; learner gender and adult reading at home.

Persistence and Spillover benefits into mathematics outcomes

The Annual National Assessments (ANA) are administered in all schools by the Department of Basic Education in September. This provides an additional data source for us to use – in particular, it offers information on mathematics outcomes and on outcomes in other grades in the school. However, the data quality is not expected to be as high as that collected by our service provider. Noisy data would be expected to cause attenuation bias in the estimated treatment effects, but this is not expected to be correlated in any way with assignment to treatment. Therefore, any statistically significant treatment effects that we do observe will be particularly interesting, but we cannot bank on conclusively answering the following three hypotheses.

Hypothesis 7*: An improvement in literacy may benefit other learning areas, such as mathematics.* We will use ANA data for grade 4 mathematics outcomes to see whether students in treatment schools benefited relative to control schools in terms of mathematics outcomes.

Hypothesis 8*: Although the intervention targeted grade 4 teachers in a school, there may be spill over benefits through shared learning amongst teachers to other grades.* We will use ANA data for literacy in grades 1, 2, 3, 5, and 6, to see whether students in untreated grades in intervention schools improved relative to students in control schools.

Hypothesis 9*: The treatment effect for intervention schools relative to control schools may diminish over time or it may grow through continued use of the new materials and pedagogies.* We will use grade 4 literacy data from the ANA, which are to be administered in September 2014 to see whether students in treated schools perform better than those in control schools (using our baseline literacy test score as a control).

F. Estimation strategy

Our realized sample in the baseline assessment consisted of 1113 students in 40 treatment schools and 1523 students in 56 control schools. One control school is completely missing in the baseline but will be included in the endline survey. We will include a dummy variable to indicate missing data on baseline, as discussed below. For the main models we do not use data from the 3 control schools that were added by the district office.

Estimation of Treatment Effects

For each of the outcomes (outlined in the Hypothesis section) for which we have collected baseline data we will start with a simple difference-in-difference calculation as follows:

$$DID = \left(\bar{y}_{T,t=2} - \bar{y}_{T,t=1}\right) - \left(\bar{y}_{C,t=2} - \bar{y}_{C,t=1}\right)$$

Our main estimation strategy will use the following ANCOVA specification:

$$Y_{i,j,t=1} = \beta_0 + \beta_1 T_j + \beta_2 S_j + \beta_3 X_j + \beta_4 Z_i + \beta_5 Y_{i,j,t=0} + \beta_6 M_{j,t=0} + \varepsilon_{i,j}$$

Where $Y_{i,j,t=1}$ is the post-treatment outcome variable of the individual learner *j* in school *i*, $Y_{i,j,t=0}$ is the baseline value of the individual learner *j* in school *i*, S_i is a vector of randomization strata dummy variables, X_j is a vector of time invariant individual characteristics (gender, age), Z_i is a vector of time invariant school and teacher characteristics (teacher qualification, teacher gender, teacher age and class size) and $M_{j,t=0}$ is a dummy indicating missing data in the baseline. T_j indicates assignment of the school to the RCUP intervention which allows us to interpret β_1 as the Intent-to-Treat effect on the schools that were assigned to the RCUP intervention. Since students are clustered in schools, standard errors will be adjusted for clustering at the school level.

In cases where we do not have baseline measures for an outcome variable (which can only occur for some intermediate outcomes where questions were added to the endline instrument) the same specification will be used but without the control for baseline measure.

The RCUP intervention is targeted at teachers in assigned treatment schools. As not all teachers might be willing to participate (or are unwilling to prepare the lessons following the lesson plans as outlined in the RCUP intervention), within treatment schools there should be some variation in the extent to which lessons were delivered and content was covered by teachers. We also have information on the number of visits from coaches. We will therefore redefine the treatment variable so as to reflect differing categories of dosage and will estimate the following equation:

$Y_{i,j,t=1} = \beta_0 + \beta_1 D_j + \beta_2 S_j + \beta_3 X_j + \beta_4 Z_i + \beta_5 Y_{i,j,t=0} + \beta_6 M_j + \varepsilon_{i,j}$ where D_j refers to the level of dosage (or coverage by teachers). We instrument D with assignment to treatment which allows us to interpret β1 as the treatment-on-the-treated effect.

Estimation of Heterogeneous Treatment Effects

Heterogeneous treatment effects will be estimated by interacting the treatment variable with the relevant variable of interest.

Unpacking causal mechanisms

If we do find that treatment assignment significantly predicts an intermediate outcome (e.g. learners taking books home) then we may attempt to predict literacy scores using the intermediate outcome variable as an explanatory variable and instrumenting it with treatment assignment. In other words, this will identify that part of the impact of taking books home that is caused by being a treatment school. We are unlikely to have high power to identify significant effects with this strategy, but if we do pick up strong effects, we may be able to shed some light on what the causal mechanisms of treatment impact were.

G. Robustness checks

1. Checking for contamination:

In order to assess whether contamination of the control group may have occurred, we exclude control schools when the distance to the nearest treatment school is less than a certain threshold. This is because it is more likely that schools close to treatment schools would be contaminated.

2. *Missing Teacher- and Learner-Level Data*

No imputation of missing data will be performed. Missing data in the baseline survey will be indicated by including dummy variables for missing data for each covariate with the value one being assigned when the value is missing and zero when the information is not missing (see above specification). Missing values in each original variable will then be re-coded as zero.

Nevertheless, we will determine correlations between assignment to treatment and the probability of missing data for teacher and learner characteristics. The results of these tests will be reported in a statistical appendix and noted in the text.

3. Attrition

We first test if treatment status significantly predicts attrition of schools or learners; and test if the missing schools/learners' performance is significantly different between treatment and control groups, based on baseline test scores.

If $Attrit_i$ indicates the probability of a school to attrite, then:

$$Attrit_i = \beta_0 + \beta_1 T_j + \beta_2 S_j + \beta_4 Z_i + \varepsilon_{i,j}$$

Where, as before, S_j is a vector of randomization strata dummy variables and Z_i is a vector of time invariant school characteristics. β_1 indicates if attrition at the school level is correlated with assignment. However, given that the RCUP study was sanctioned by the district officials of the Department of Basic Education, it is unlikely that entire schools attrite. Nevertheless, if attrition is to happen, we assume that control schools are more likely to attrite especially given that the post-test is administered in the final week of the term.

We also expect some pupils to attrite. Thus, if $Attrit_{i,j}$ indicates the probability of a learner in a particular school not to write the post-treatment test, then:

$$Attrit_{i,j} = \beta_0 + \beta_1 T_j + \beta_2 Y_{i,j,t=0} + \beta_3 T_j * Y_{i,j,t=0} + \beta_4 X_j + \varepsilon_{i,j}$$

Where β_1 shows the effect of being in a treated school on the likelihood of attriting, β_2 indicates the influence of baseline achievement on not writing the post-test, and β_3 shows the interaction effect on attrition of being a learner in a treatment school given performance in the pre-test.

As a further robustness check if we do find that treatment assignment significantly predicts attrition (at the 90% level of confidence), we will conduct the following bounding exercise similar to that proposed by Lee (2009):

We will observe that X% are missing in the treatment group and Y% in the control group with X<Y);
Y-X can now be regarded as "excess missing" cases in the control group;
Randomly select Y-X number of treatment group students;
Delete the endline score for these students;
Impute a value of Z+G as the endline score for these students (where Z is their baseline score and G is the gain score experienced at the 10th percentile of control group students);
Run the analysis using the a) observed values and b) imputed observations, and analyse sensitivity of results.
Redo with G equal to 25th percentile and G equal to 50th percentile.

4. Interference with randomization

We also have a problem posed by the replacement of 3 control schools. These schools were replaced on the request of the district office and the reasons provided were legitimate and would have applied equally to treatment schools had it been necessary. This means that the remaining 57 control schools still serve as a valid comparison group to the treatment schools. For the main estimation models we thus use only these 57 control schools and do not use the 3 new control schools.

Nevertheless, it is possible that we actually lost 3 schools which would have had a higher or lower propensity to improve than the average amongst the control schools. If this were the case our remaining control group of 57 schools is no longer completely valid. In order to test the sensitivity of our main result to this possibility we conduct the following bounding exercise:

We use the baseline scores for the 3 new control schools. As a lower bound of the treatment estimate, we will impute new endline scores such that the gains (from baseline to endline) experienced by students in these 3 schools would be the same as the gain achieved at the 90^{th} percentile of gains amongst control group students. This conservatively simulates the scenario that the 3 original schools had a systematically higher propensity to improve. As an upper bound of the treatment estimate, we will impute new endline scores such that the gains (from baseline to endline) experienced by students in these 3 schools would be the same as the gain achieved at the 10^{th} percentile of gains amongst control group students. This conservatively simulates the scenario that the 3 original schools had a systematically lower propensity to improve.

We also present the results when including the actual observed endline scores for the 3 new control schools, as if they had been initially selected.

H. Cost effectiveness

In order to establish the cost effectiveness of the RCUP intervention we calculate the standard deviations gained per US$100 spent on treatment. This allows us to make comparisons with other studies reported on in Kremer, Brannen and Glennerster (2013). We use the estimated treatment effect size from the main Intent-To-Treat equation. We convert costs from Rand values to US dollars using the Rand-Dollar exchange rate as at the close of the South African markets on June 16 2014. The rate to be used is thus R10.75 to US$1.

References

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. 2010. *Pitfalls of participatory programs: Evidence from a Randomized Evaluation in Education in India.* 2:1: 1-30.

Banerjee, A., & Duflo, E. (2011). *Poor Economics.* London: Penguin Books.

Friedman, W., Gerard, F., & Ralaingita, W. (2010). International independent evaluation of the effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead"(RLL) program in Mali, mid-term report. *Research Triangle Park, NC: RTI International*.

Glewwe, P. (2002). Schools and skills in developing countries: Education policies and socioeconomic outcomes. *Journal of economic literature*, *40*(2), 436-482.

He, F., Linden, L., & MacLeod, M. (2008). How to teach English in India: Testing the relative productivity of instruction methods within the Pratham English language education program. *New York, United States: Columbia University. Mimeographed document*.

Kremer, M., Brannen, C., & Glennerster, R. 2013. The Challenge of Education and Learning in the Developing World. *Science.* 340: 297-300.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects, *Review of Economic Studies* **76**: 1071–1102.

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2013). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. *Unpublished manuscript.[58, 76]*.

McEwan, P. J. (2013). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Unpublished manuscript, Wellesley College*.

Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development*.

Taylor, S. (2011). Uncovering indicators of effective school management in South Africa using the National School Effectiveness Study. *Stellenbosch Economic Working Papers* (10/11).

APPENDIX E: CHAPTER 4 EGRS Baseline Report

THE EARLY GRADE READING STUDY:

A report on the baseline data collection and Year 1 programme activities



IMPROVING EARLY GRADE READING IN SOUTH AFRICA [P2.10.SA.IE]

BASELINE REPORT, 29 OCTOBER 2015

RESEARCH TEAM (AUTHORS)

Stephen Taylor

Jacobus Cilliers

Brahm Fleisch

Cas Prinsloo

Servaas van der Berg

Vijay Reddy

The Early Grade Reading Study team gratefully acknowledges the generous support from the donors and partners listed below.

basic education
Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

Responsible Department: Project Design, Management and Governance

unicef
unite for children

planning, monitoring and evaluation
Department:
Planning, Monitoring and Evaluation
REPUBLIC OF SOUTH AFRICA

PROUDLY SUPPORTED BY ANGLO AMERICAN

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

Project Administration

ZENEX FOUNDATION

GAUTENG PROVINCE
EDUCATION
REPUBLIC OF SOUTH AFRICA

Adapted from a programme developed and paid for by the Gauteng Department of Education.

class act

Service Provider

HSRC
Human Sciences Research Council

3ie International Initiative for Impact Evaluation

CONTENTS

Abbreviations and Acronyms 5
List of tables 6
List of figures 7
Executive summary 8
Background to the Early Grade Reading Study 11
Description of Interventions 12
Theory of change 13
Research site 17
Evaluation design 20
Treatment assignment and sample selection 20
Instrument development and piloting 22
Ethical clearance 23
Data collection 23
Data capturing and cleaning 28
Baseline Results 29
Learner test scores 29
Pupil characteristics 42
Parent characteristics 43
Teachers 47
Test performance 47
Teacher Characteristics 50
School and school principal 52
What Predicts Pupil Test Scores? 54
Pupil Characteristics 55
School and Teacher Characteristics 55
Parent Characteristics 57
Progress report on implementation of interventions 61
EGRS Treatment 1 (training) 61
EGRS Treatment 2 (coaching) 64
EGRS Treatment 3 (parents) 67
Next steps in the project 70
References 71

ABBREVIATIONS AND ACRONYMS

DBE	Department of Basic Education
PIRLS	Progress in International Reading Literacy Study
RCT	Randomized Controlled Trial
ANA	Annual National Assessments
EGRS	Early Grade Reading Study
SGB	School Governing Body
RDD	Regression Discontinuity Design
EGRA	Early Grade Reading Assessment
HSRC	Human Sciences Research Council
CRC	Community Reading Coach
GPLMS	Gauteng Primary Language and Mathematics Strategy

LIST OF TABLES

Table 1: Number of schools by phase in Dr Kenneth Kaunda and Ngaka Modiri Molema _____ *18*
Table 2: Grade 3 learners achieving 50% and above by subject _____ *19*
Table 3: Grade 6 learners achieving 50% and above by subject _____ *19*
Table 4: Number of learners successfully tested per school _____ *25*
Table 5: Numbers of returned learner tests and parent questionnaires _____ *26*
Table 6: Number of parent questionnaires returned per school _____ *26*
Table 7: Numbers of returned teacher and principal questionnaires _____ *27*
Table 8: Summary statistics for items in sub-tests A and C _____ *30*
Table 9: Summary statistics for items in sub-test D _____ *31*
Table 10: Summary statistics – Aggregate test scores _____ *31*
Table 11: Cronbach's alpha for Section A (expressive vocabulary) _____ *32*
Table 12: Cronbach's alpha for Section C.1 (short-term memory – words) _____ *34*
Table 13: Cronbach's alpha for Section C.2 (short-term memory – numbers) _____ *34*
Table 14: Cronbach's alpha for Section D (Phonological Awareness) _____ *36*
Table 15: Balance tests _____ *42*
Table 16: Descriptive statistics – learner age and gender _____ *42*
Table 17: Selected descriptive statistics – teacher characteristics _____ *52*
Table 18: Selected descriptive statistics – school principal questionnaire _____ *53*
Table 19: Performance by district, learner age and gender (OLS Regressions) _____ *55*
Table 20: Performance by school principal, location, and socio-economic background _____ *56*
Table 21: Performance by teacher characteristics _____ *57*
Table 22: Parent characteristics and learner performance _____ *58*
Table 23: Attendance at Treatment 1 training events _____ *62*
Table 24: Structure of treatment 2 _____ *64*
Table 24: Attendance at training events _____ *65*
Table 25: Attendance at CRC training events _____ *68*
Table 26: Attendance at parent training events _____ *68*

LIST OF FIGURES

Figure 1: Theoretical diagram of how reading acquisition occurs ____ 14
Figure 2: Map of South Africa showing education districts ____ 17
Figure 3: Highest Education level for adults aged 20 and older ____ 19
Figure 4: Diagram showing sampling procedure ____ 21
Figure 5: Map of North West province showing schools by treatment assignment ____ 22
Figure 6: Kernel Density Curves for Section A (expressive vocabulary) by treatment arm ____ 32
Figure 7: Kernel Density Curves for Section B (letters correct) by treatment arm ____ 33
Figure 8: Kernel Density Curves for Section C (short-term memory) by treatment arm ____ 35
Figure 9: Kernel Density Curves for Section D (Phonological Awareness) by treatment arm ____ 36
Figure 10: Kernel Density Curves for Section E (word recognition) by treatment arm ____ 37
Figure 11: Number of words correct in Section F (sentence reading) ____ 38
Figure 12: Percentage of learners scoring 0, 1, 2 and 3 for Section F comprehension ____ 38
Figure 13: Relationship between words read and comprehension ____ 39
Figure 14: Item Information Functions from a 1-parameter IRT model ____ 40
Figure 15: Test Information Function from a 1-parameter IRT model ____ 40
Figure 16: Kernel Density Curves for composite test score by treatment arm ____ 41
Figure 17: Relationship to pupil ____ 43
Figure 18: Education of Guardian ____ 44
Figure 19: How often do you read to your child? ____ 45
Figure 20: Number of books at home ____ 45
Figure 21: Number of hours read for own pleasure ____ 46
Figure 22: Check that child is doing homework ____ 46
Figure 23: Agreement with the following statement: "Our school's learners read very poorly." ____ 47
Figure 24: Words read by teachers in 60 seconds ____ 48
Figure 25: Scores on the teacher comprehension test ____ 49
Figure 26: Relationship between words read and questions answered correctly ____ 49
Figure 27: Teacher qualifications ____ 50
Figure 28: At what grade should pupils be able to read fluently? ____ 51
Figure 29: What proportion of children will be able to read by grade 3? ____ 51
Figure 30: School principal highest level of education ____ 53
Figure 31: School principal's estimate of average level of parent education ____ 53
Figure 32: Proportion of learners with both parents employed according to school principal ____ 54
Figure 33: Location of the school according to school principal ____ 54
Figure 34: Parent education and learner test scores ____ 59
Figure 35: Parent reading to child and learner test scores ____ 59
Figure 36: Parent checking homework and learner test scores ____ 60
Figure 37: Parent's perceived responsibility for learning and learner test scores ____ 60

EXECUTIVE SUMMARY

This report describes the initial phase of the South African Department of Basic Education's Early Grade Reading Study (EGRS), especially the results from the baseline data collection. The EGRS project involves the implementation and evaluation of three alternative programmes all aimed at improving the acquisition of home language reading and literacy. The project is being implemented in two districts in the North West Province, in which the main home language is Setswana. The EGRS is working with the grade 1 class of 2015 for a two-year period, following the same learners into grade 2 in 2016.

The first intervention (implemented in 50 schools) provides teachers with lesson plans, additional reading support materials and training at centralized workshops twice a year. The second intervention (implemented in a different group of 50 schools) provides teachers with the same set of lesson plans and additional reading support materials but provides ongoing support to teachers through monthly on-site coaching and small cluster training sessions. The third intervention (implemented in a further 50 schools) holds weekly meetings with grade 1 parents to inform them of the importance of learning to read in the early grades and to empower them with knowledge and tools to become involved in their own child's reading acquisition.

Assignment to each of the three intervention or "treatment" groups and to a further group of 80 control schools was done through a computerized lottery. This ensures comparability across the groups. This randomized assignment is the key design feature of the EGRS and will be the basis for making claims about the causal impacts of each intervention on reading outcomes, when measured at the end of grade 1 and again at the end of grade 2.

The baseline data collection was administered by the Human Sciences Research Council (HSRC) in all 230 schools in February 2015. A random sample of 20 grade 1 learners per school participated in oral assessments of reading and pre-reading skills. Questionnaires were also administered to the school principal, to all grade 1 teachers and to parents of the 20 tested learners.

The learner tests were adapted from the well-known Early Grade Reading Assessment (EGRA) tool and covered the following skills in Setswana: expressive vocabulary, letter recognition fluency, short-term memory, phonological awareness, word recognition fluency, sentence reading and sentence comprehension.

The baseline testing confirms the success of the randomization: on all measures of reading ability there is a good balance across the four treatment groups. Some test subtasks yielded ceiling effects (where many learners achieved the maximum score) and other subtasks yielded floor effects (where many learners achieved the minimum score) but across the entire test there was a good variation in scores.

Girls outperformed boys on the reading tests. This advantage for girls is consistent with what is observed in standardized tests for higher grades in South Africa, such as in the Annual National Assessments. It is interesting that this gender gap is evident right at the start of grade 1, which would suggest that the girl advantage may be due to some factor other than school practices, most likely differences in the physiological development of girls and boys at this age.

The parent questionnaires indicate that the majority of parents or guardians in participating schools have low levels of education and they also have high levels of unemployment. Parent education is also predictive of cognitive ability and basic reading skills at the start of school. Similarly, the children of parents/guardians who read with them performed better on the baseline tests.

The majority of schools in the sample are poor and situated in rural areas. Children in rural schools performed worse in the baseline assessment than those in urban or semi-urban areas. It was interesting that learner performance was significantly better in the district of Ngaka Modiri Molema than in Dr Kenneth Kaunda. This was not expected and will therefore be closely monitored as the project continues.

The grade 1 teachers are almost always female, and are rather old with an average age of 50. About 26% of teachers are 56 years old or older. Less than 10% of teachers are younger than 40. This has implications for the future provisioning of Foundation Phase education in general and for home language learning in particular (since other evidence suggests that low proportions of new Foundation Phase teacher graduates are specialized in the African languages). It may also affect the theory of change for the interventions if there are any differences in the way older teachers versus younger teachers react to support programmes. One teacher characteristic that was positively correlated with child learning outcomes was the teacher's own reading comprehension. It is unlikely that this reflects a causal relationship between teacher quality and learner performance since learners have just joined the school. It is possible that this reflects a selection effect where both learners and teachers select themselves into better schools. If this is indeed the case it represents a striking phenomenon: stronger children tend to be taught by stronger teachers.

Throughout the course of 2015 the three interventions have been running in schools. The lesson plans and reading support materials have been delivered to all schools in Treatment groups 1 and 2. Two training sessions have been held for teachers in Treatment 1 schools to date, in February and in July 2015. These 2-day training sessions were well attended – 100% attendance at the first and 85% attendance at the second. Schools in Treatment 2 have been receiving monthly coaching visits and afternoon clustered coaching sessions – three coaches share the 50 schools more or less equally between them. Teacher attendance rates at the cluster sessions were 100% in Term 2, 82% in Term 3 and 93% in Term 4. Some qualitative analysis based on classroom observation in a handful of Treatment 1 and 2 schools would suggest that the implementation of the prescribed lesson plans has been done with greater

fidelity in Treatment 2 schools (coaching) than in Treatment 1 schools (training). However, this is based on a small sample of schools so no conclusions can be drawn with certainty.

One Community Reading Coach (CRC) has been recruited per Treatment 3 school and trained to run weekly afternoon sessions open to all grade 1 parents. A total of 30 sessions is scheduled for each year covering a total of 10 topics per year. Each topic has 3 sessions where the topic is the same but the activities of the session differ. Thus a parent can attend roughly 1 in 3 sessions and still be exposed to all topics, while parents who attend more regularly can still enjoy fresh activities. In a few schools it has proven difficult to recruit a CRC or the CRC has had to be replaced. Parent attendance has also been a challenge in these schools with attendance rates dropping from 35% for the orientation sessions and Topic 1 to 18% for Topic 4. Creative ways to encourage greater attendance in 2016 will need to be considered.

Interventions are scheduled to continue throughout 2016. A midline data collection is taking place between the 26th of October and the 13th of November 2015. The endline data collection is scheduled for October/November 2016.

BACKGROUND TO THE EARLY GRADE READING STUDY

The acquisition of reading is foundational to all subsequent learning; yet the majority of South African children are being left behind in this regard. The PIRLS study of 2006 showed that a striking 80% of South African children were not yet reading with comprehension after five years of schooling. The problem is particularly severe amongst poor children. Consequently, massive inequalities in educational achievement are established early in primary school and there is no evidence of these inequalities being reduced in later years. Therefore, early interventions, such as improving the acquisition of reading amongst poor children, can be expected to have larger effects than interventions later in the school programme.

The recently introduced Annual National Assessments (ANA) have raised public awareness of the weak literacy achievement of children in the primary school grades. Although the DBE and provincial education departments are implementing various strategies to support early grade reading, there is little or no sense of what is working and why. Moreover, there are competing models of support in the system. Some provinces favour the traditional model of teacher training workshops, while the province of Gauteng has provided additional graded readers and clearly scripted lesson plans together with specialist reading coaches who visit teachers on monthly basis to observe lessons and offer assistance. It is important that a national reading strategy be based on scientific evidence regarding what most improves the acquisition of reading.

A randomized controlled trial (RCT) design allows a credible estimation of the true causal impact of interventions, and thus has the potential to inform responsible policy decisions. Through the use of a lottery to allocate schools to intervention and control groups it is possible to construct a credible "counterfactual" scenario – what would have happened to those who received an intervention had they not received that intervention.

Moreover, by directly comparing the impacts on reading outcomes of alternative programmes, each with different cost implications, we can identify the most cost-effective intervention. This project is designed to explicitly compare the impact and cost of a new model of teacher development (on-school support) to the impact and cost of a more traditional model (training at central venues). The third intervention, which aims at improving parent involvement in schools and in home-based reading activities, relies on a rather different theory of change and is less expensive. By measuring the success of each intervention on the same scale, this project will provide a sense of the cost-effectiveness of different policy options.

The primary implementing partner is the South African government, in particular the Department of Basic Education. A key role is also being played by the North West provincial education department, which is contributing financially and is championing the project within the schools.

A service provider has been appointed to run the three interventions on behalf of the DBE for the purposes of this impact evaluation. The service provider is an organisation called "Class Act", which is highly involved in partnerships with government to run literacy interventions. For

example, “Class Act” was a service provider in the Gauteng Province’s implementation of the Gauteng Primary Literacy and Maths Strategy (GPLMS) over the last few years. Programme interventions are being funded by a coalition of donors, including the ZENEX Foundation, UNICEF, Anglo American and the Department of Planning, Monitoring and Evaluation in the Presidency. These funds are being managed by the University of the Witwatersrand, which ran a tender for the service provider work and subsequently entered into a contract with Class Act.

The evaluation side of the project is being supervised by the Research Team while the data collection and capturing is being managed by South Africa’s Human Sciences Research Council (HSRC). The evaluation is being funded by the International Initiative for Impact Evaluation (3ie).

DESCRIPTION OF INTERVENTIONS

This study evaluates three different interventions, all aimed at improving early-grade reading in the home language, which in the case of the North West province is Setswana. All three interventions work with children entering Grade 1 at the start of 2015 over a two-year period (thus working with grade 2 learners in 2016).

Treatment 1: Training, scripted lessons, graded readers.

Treatments 1 and 2 aim to apply the same set of instructional practices in the teaching of home language literacy in grade 1 and 2 classrooms. Both treatments provide teachers with lesson plans, which are aligned to the curriculum as specified in the Curriculum and Assessment Policy Statements (CAPS) for home language literacy in the Foundation Phase. The lesson plans provide detailed specification for each lesson including information on methodology and content to be taught for each instructional day. The lesson plans incorporate the use of learning support materials including the government-provided workbooks as well as certain additional materials (graded reading booklets, flash cards, posters, etc.), which are provided through the EGRS. The graded reading booklets provide a key resource for the teacher to use in group-guided reading and individual work so as to facilitate reading practice at an appropriate pace and sequence of progression.

Treatment 1 trains the teachers on how to use the lesson plans and accompanying materials through central training sessions, each lasting 2 days, and occurring twice yearly. The first session was conducted in February 2015 and the second occurred in July 2015. Similar sessions are scheduled for 2016.

Treatment 2: Reading Coaches, scripted lessons, graded readers.

Exactly the same set of instructional materials (scripted lesson plans, graded reading booklets and other materials) is provided to Treatment 2 schools. However, instead of central training

sessions, ongoing support to teachers consisting of regular (monthly) on-school coaching from specialist "reading coaches" is provided. In addition to these on-site visits, there are occasional meetings with the coach and a small cluster of nearby Treatment 2 schools. The evaluation of treatments 1 and 2 should thus shed light on whether the fairly prescriptive instructional regime has the ability to improve reading acquisition and whether the mode of teacher support is important in mediating effectiveness.

Treatment 3: Parental involvement

Treatment 3 is designed to promote parental involvement to support their children's reading progress. At each of the 50 schools in this treatment arm a Community Reading Coach (CRC) was recruited. The CRC was identified through communication with the school principal who recommended a suitably qualified but available person in the community. The CRCs attend a 1-day training session facilitated by the service provider (Class Act) at the start of each school term (quarterly). The CRCs are trained to deliver weekly training sessions for grade 1 parents at their respective schools. A total of 30 sessions is scheduled for each year covering a total of 10 topics per year. Each topic has 3 sessions where the topic is the same but the activities of the session differ. Thus a parent can attend roughly 1 in 3 sessions and still be exposed to all topics, while parents who attend more regularly can still enjoy fresh activities. For their services, CRCs are paid a stipend of R400 per month (about $35).

The topics covered in these sessions include the importance of learning to read for later educational and labour market success, training on how to support their child's reading at home and the provision of low-cost materials and reading games to use at home.

THEORY OF CHANGE

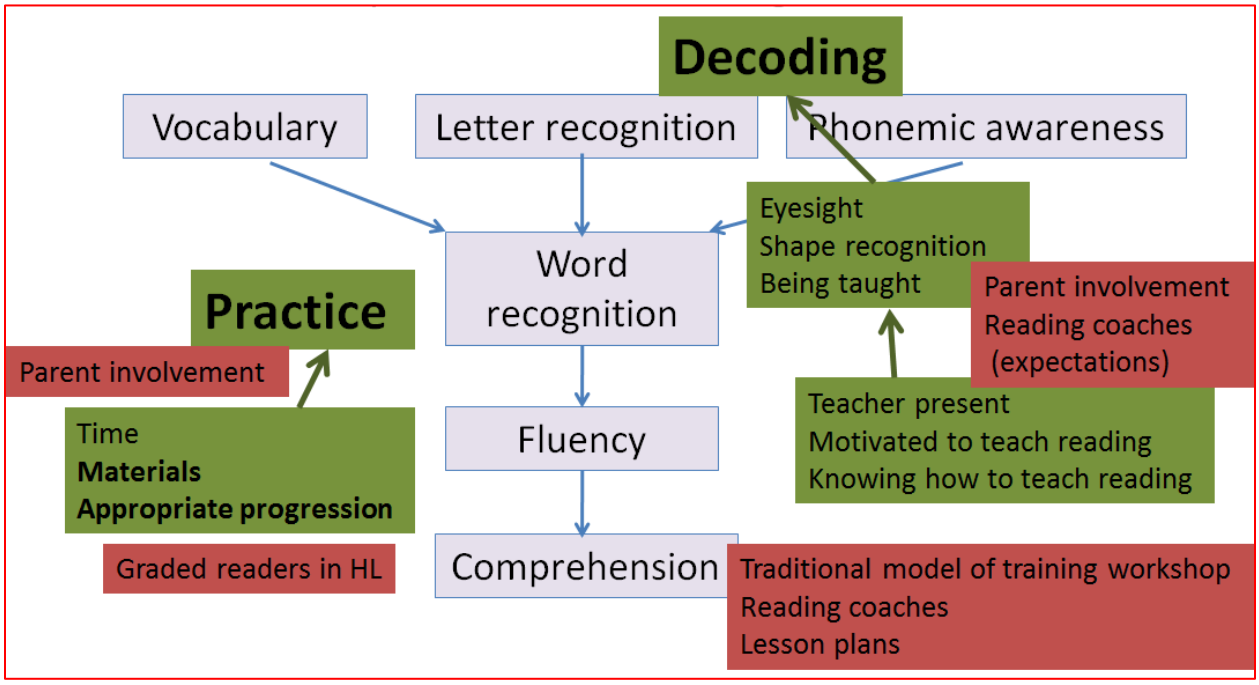
Reading acquisition

All three interventions relate to the educational theory of how reading acquisition occurs. Reading comprehension is the product of two components: vocabulary and decoding. To a great extent vocabulary (and more broadly language acquisition) comes naturally through speaking and hearing others speaking. Through speaking and hearing others speaking, phonological awareness also develops - this involves sound segmentation and recall of sound patterns. This phonological awareness is important for children to learn to decode. Particular written shapes are associated with particular sounds. Decoding thus consists of letter recognition and phonemic awareness. Unlike learning to speak, decoding does not come naturally; it is a method that must be taught systematically. It is important to emphasize that reading is produced by the product of vocabulary and decoding: If one has a perfect vocabulary but has not been taught the method of decoding one will not be able to read at all. Letter recognition and phonemic awareness are mastered through systematic teaching and consistent practice. This leads to the next stage of reading acquisition: word recognition. Through practice

and appropriate progression from simpler sounds and words to more complex ones word recognition becomes established leading to the next phase of reading acquisition: fluency. It is only once decoding and word recognition have become fluent that it is possible to reach the ultimate goal of reading comprehension.

In order to learn the basics of decoding, a child requires a teacher who is present, capable and motivated to deliver systematic reading instruction. In order for decoding to become fluent a child requires suitable graded materials and the discipline (perhaps imposed) to practice a lot. The interventions to be tested in this study address these needs in various ways. Figure 1 presents a theoretical diagram illustrating how reading acquisition occurs, what supportive conditions need to be in place and how each of the interventions being evaluated in the EGRS address key points in the development of reading acquisition.

Figure 1: Theoretical diagram of how reading acquisition occurs



There is a growing body of evidence from developing countries that early grade reading interventions can have a significant impact. The "EGRA Plus" programme administered in Liberia produced substantial gains in reading achievement relative to comparison children who did not receive the programme. Key aspects of this programme included a cascading model of reading coaches, the distribution of scripted lesson plans and reading assessment tools, and the dissemination of report cards to parents (Gove and Wetterberg, 2011). A supplementary reading curriculum administered in India also produced significant improvements in both public schools and pre-schools (He, Linden and MacLeod, 2009).

However, these studies cannot tell us which component of the intervention is responsible for the success of the program. This is important for policy purposes, because we want to find the most cost-effective intervention which could be scaled up by government. For example, the "EGRA plus" programme in Liberia was clearly highly resource-intensive because it required ongoing monitoring from qualified reading coaches, but we do not know if one might be able to reach the same results with a sub-component of the program. Moreover, there is uncertainty about the transferability of the findings given different language and social contexts.

Similar programs have been implemented in South Africa, but since they were not credibly evaluated, we do not know if they truly improved pupils' reading acquisition. The Department of Basic Education typically holds training programs similar to our intervention 1; and Gauteng has implemented a model of reading coaches, similar to intervention 2. Since it has not been possible to produce a robust empirical impact evaluation of these programmes, we do not know if they truly work or not. Fleisch and Schoer (2014) attempted a Regression Discontinuity Design (RDD) to evaluate the impact of the Gauteng Primary Language and Mathematics Strategy (GPLMS) and findings pointed to a positive impact, though the findings were tentatively made given significant data constraints. Sailors et al (2010) evaluated a reading intervention in South Africa, which followed a similar model to intervention 2, but there are large methodological challenges to the study.

There is also a growing international literature providing information to parents and fostering parental involvement in schools can improve learning outcomes, but there is much we still do not know. In Pakistan, pupils who came from villages where the community was provided with information of school performance performed better in independently administered tests, compared to pupils from villages where no such information was administered. The improvement was particularly large for schools with low initial learning outcomes (Andrabi et al, 2013). In a different programme in India, school communities were informed of their school performance and also educated on their rights, roles and responsibilities in school governance through 8 public meetings. Education performance improved as a result (Pandey et al, 2013). However, in a recent impact evaluation in Kenya, informing parents on their child's reading progress had zero impact (Lieberman, Posner and Tsai, 2013). The authors hypothesize necessary conditions for an information-intervention to work, all of which we address in our study: (i) information is new; (ii) it highlights under-performance and potential to improve; (iii) it is combined with measures which enable parents to act on this information.

All interventions aim to improve reading acquisition in the home language. Strictly speaking, the targeted outcome is home language literacy more broadly, since this is the Foundation Phase curriculum area being given support through our programmes. The choice to address home language literacy is motivated by research showing long-term benefits to strong home language skills prior to switching to a second language. Taylor and Coetzee (2013), for instance, show that in South Africa using home language as the language of instruction during grades 1, 2 and 3 has been associated with better English acquisition in grades 4, 5 and 6.

Intervention 1:

This programme is intended to impart the capacity to ensure that it is possible for the teacher to provide effective and systematic reading instruction in the classroom. Scripted lessons provide a structure to assure systematic practice and learning based on sound pedagogical theory. It can act as a substitute to low teacher capability or low motivation to prepare lesson plans. The accompanying reading materials aim to ensure that all the necessary instructional infrastructure is in place for a systematic reading programme to be effectively implemented.

Intervention 2:

The reading coach intervention provides more intensive training to improve teacher capacity. The assumption is that, just like learning to read, the ability to teach is a skill that needs to be developed over time and might not be accomplished in one-off training. Furthermore, the reading coaches could also improve teacher motivation as they are frequently monitored, provided with much-needed additional support, and can also find inspiration from watching an excellent example provided occasionally by coaches. This programme thus addresses both teacher capacity and teacher motivation. Another way to describe the difference between Treatments 1 and 2 is that while they share an underlying pedagogical theory of change (centered around instructional alignment and coherence using prescriptiveness as a vehicle), they differ in their theory of action (where Treatment 2 has a stronger component focused on changing behavior using accountability and motivation).

Intervention 3:

Parents pay a critical component to learning to read, as it requires continuous practice, both at school and at home. For parents to be *willing* to play this role they need to appreciate (i) the importance of reading; and (ii) that their child is most likely not learning enough at school and requires additional support. This is the purpose of the information. For parents to be *able* to play this role, they need to understand the necessary steps in learning to read and also have appropriate material to practice reading with their child. This is the purpose of the training and additional practice material.

Each of these three interventions has a different theory of change and also has different cost implications. Treatment 3 has the lowest cost amounting to approximately R16 000 per school per year (i.e. about $1200). Treatment 1 costs approximately R34 000 per school per year (i.e. about $2600). Treatment 2 is the most costly, amounting to approximately R63 000 per school per year (i.e. about $4800).

RESEARCH SITE

The EGRS is being implemented in the North West province, in the districts of Dr Kenneth Kaunda and Ngaka Modiri Molema. The North West province was chosen on the basis of 1) it being a relatively poor province, thus making it relevant to the majority of the underperforming South African school system; 2) it is relatively homogenous in terms of home language (Setswana) making it more affordable to develop learning support materials in a single language; 3) it is within driving distance from the Gauteng province where the national DBE is located; and 4) the senior management of the North West provincial education department were eager to partner with the DBE on this project. The district of Bojanala was excluded because another special targeted intervention was taking place in that district at the same time. The district of Dr Ruth Segomotsi Mompati was excluded since it is particularly far West of Gauteng and since enough schools existed in the districts of Dr Kenneth Kaunda and Ngaka Modiri Molema. Figure 2 shows a map of South Africa divided into the 83 education districts.

Figure 2: Map of South Africa showing education districts

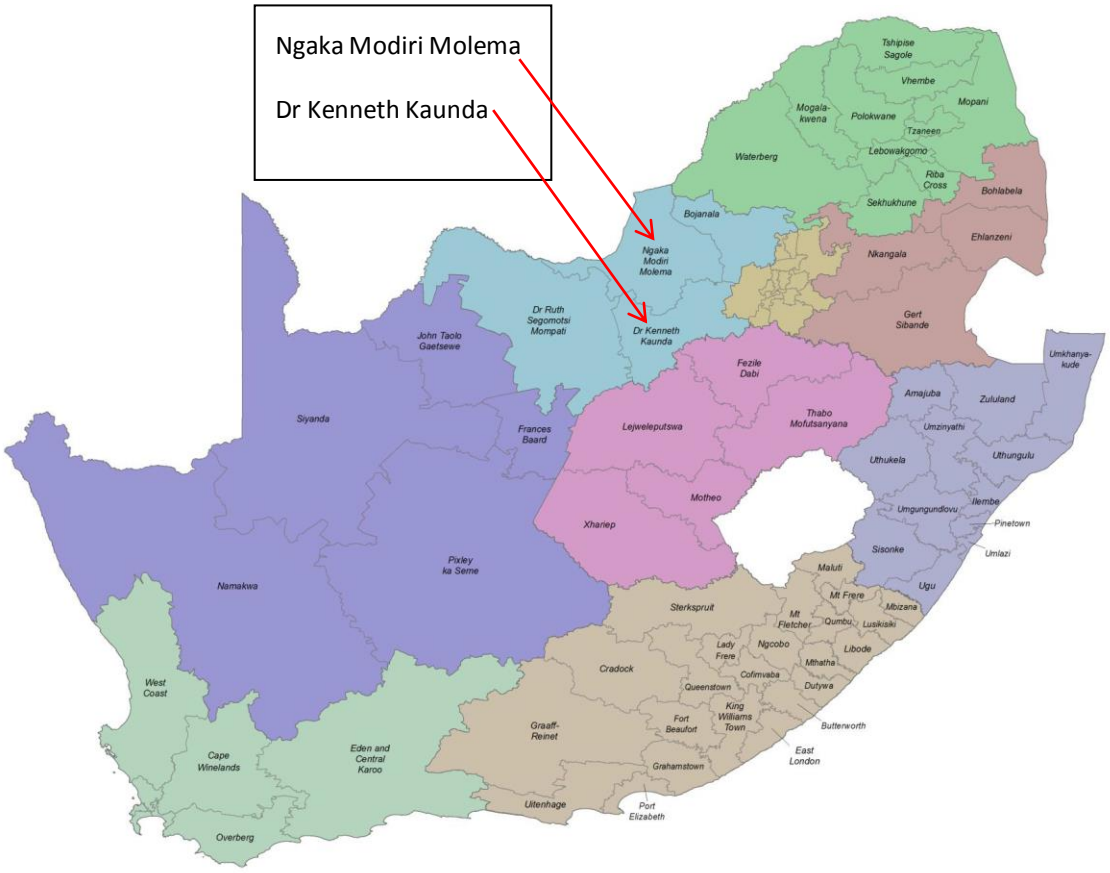


Table 1 below shows the total number of ordinary schools by phase for both Dr Kenneth Kaunda and Ngaka Modiri Molema districts in 2014. We see that Ngaka Modiri Molema district has the highest number of schools across all categories. Of the 248 schools in Dr Kenneth

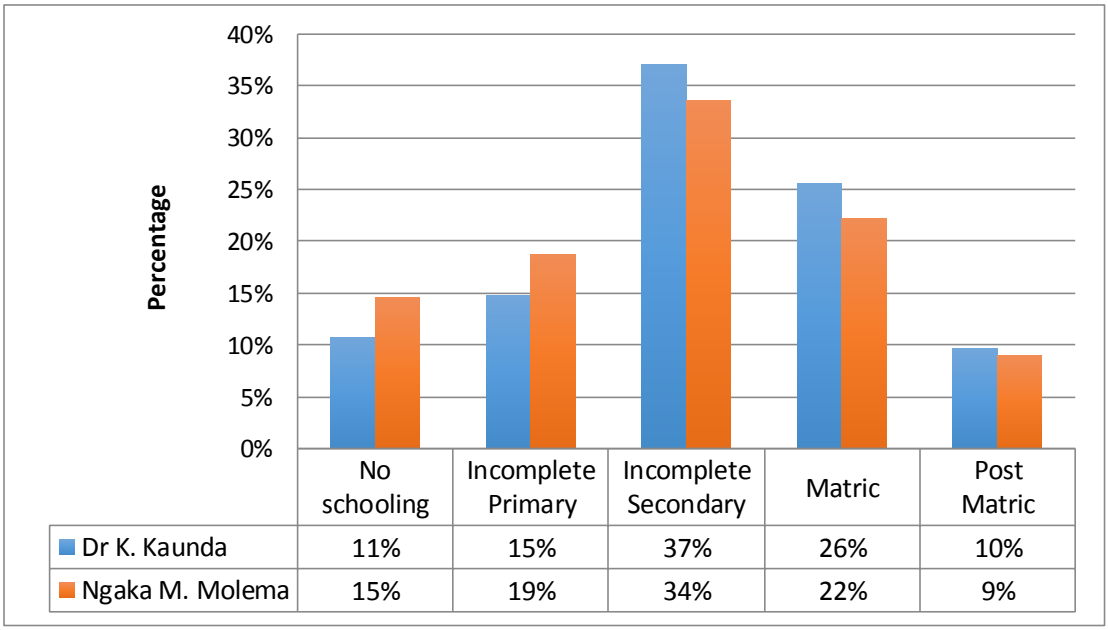
Kaunda district, 14 are independent schools while 11 of the 404 schools in Ngaka Modiri Molema district are independent schools. In Dr Kenneth Kaunda, 81% of schools are no-fee schools (classified as Quintile 1, 2, and 3 according to the official school poverty classification) while the equivalent figure was 91% of schools in Ngaka Modiri Molema district. This confirms that these two districts are largely poor and rural parts of South Africa. The choice of these areas for the EGRS project was deliberate so as to optimize the relevance of the study's findings to the large, underperforming and poor sections of South Africa's school system.

Table 1: Number of schools by phase in Dr Kenneth Kaunda and Ngaka Modiri Molema

	Dr Kenneth Kaunda		Ngaka Modiri Molema	
	Number	%	Number	%
Primary	149	60%	247	61%
Secondary	54	22%	76	19%
Combined	42	17%	67	17%
Intermediate	3	1%	14	3%
Total	248	100%	404	100%

In the 2011 Census, people were asked to indicate the highest level of education that they had completed. It referred to the highest level completed, not the level currently in, if the person was still studying. Figure 3 shows the education levels of adults aged 20 and older by district. The category 'Matric' refers to the secondary school leaving examination. This figure shows that Dr Kenneth Kaunda district had higher proportions of people who had a matric and post matric qualifications compared to those in Ngaka Modiri Molema district. Overall, this figure implies that the majority of people who would be parents to Grade 1 pupils would have relatively low levels of education.

Figure 3: Highest Education level for adults aged 20 and older



The Annual National Assessment (ANA) results provide an indication of school performance at the primary school level. It should be noted, however, that results are not comparable across time or across subjects or grades, since the tests cannot be equated to each other. In 2012 Dr Kenneth Kaunda performed better than Ngaka Modiri Molema. However, the opposite was true in 2013. This seems strange, and may reflect differential test administration and marking practices across time and district. The broad point to note is that language and mathematics performance in both of these districts is at a low level, allowing much room for improvement.

Table 2: Grade 3 learners achieving 50% and above by subject

Subject	Year	Dr Kenneth Kaunda	Ngaka Modiri Molema
Mathematics	2012	30%	18%
	2013	49%	48%
Language	2012	53%	41%
	2013	44%	49%

Table 3: Grade 6 learners achieving 50% and above by subject

Subject	Year	Dr Kenneth Kaunda	Ngaka Modiri Molema
Mathematics	2012	9%	7%
	2013	15%	23%
Language	2012	25%	19%
	2013	40%	45%

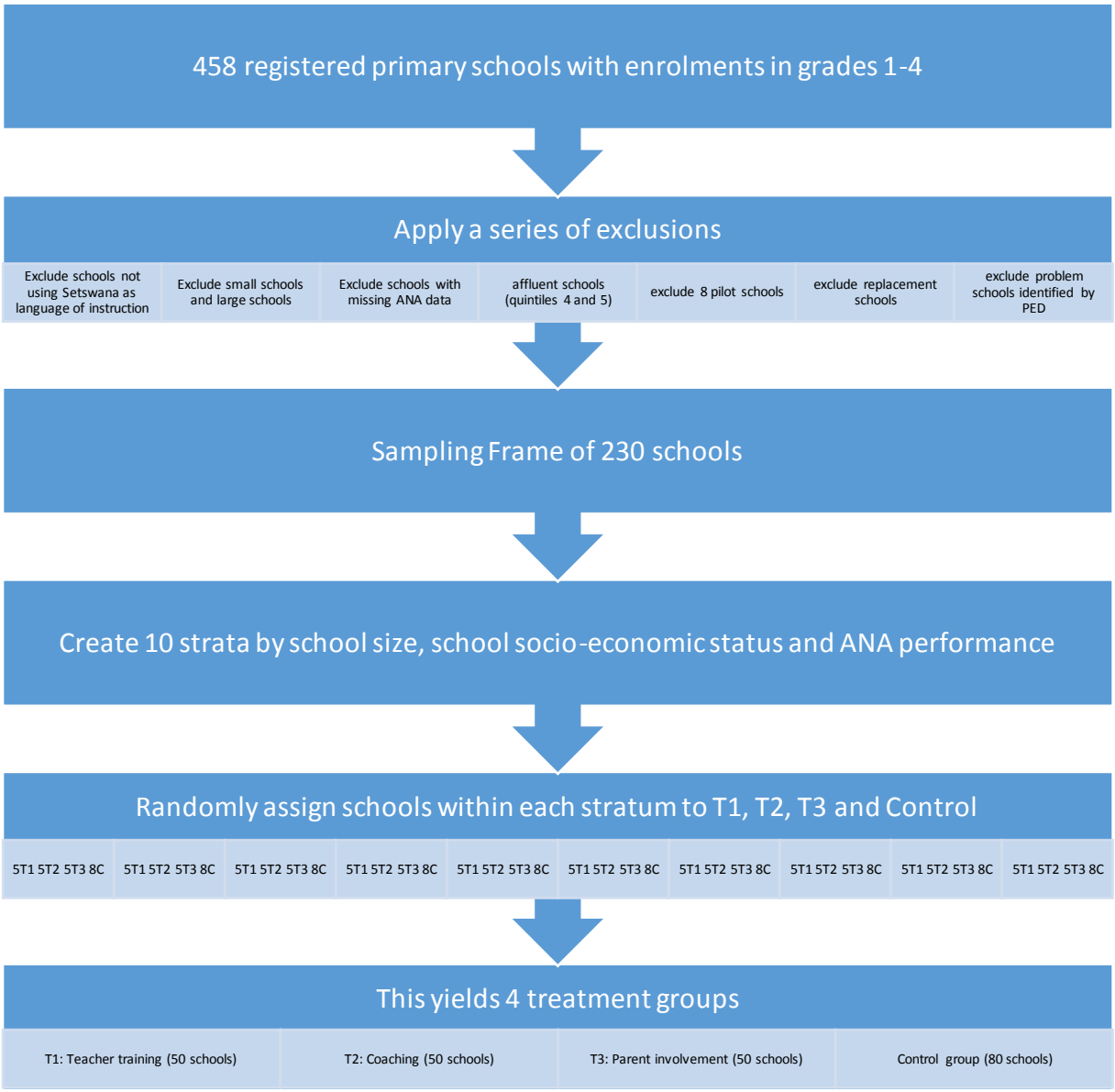
EVALUATION DESIGN

TREATMENT ASSIGNMENT AND SAMPLE SELECTION

Through a process of elimination we developed a sampling frame of 230 eligible schools. Beginning with 458 primary schools registered in 2014 administrative data in the districts of Dr Kenneth Kaunda and Ngaka Modiri Molema we started by excluding relatively affluent schools (those in quintiles 4 and 5). Next, we excluded schools in which the language of instruction in the Foundation Phase was not Setswana. We excluded schools which were missing in the 2014 ANA dataset. We also excluded 8 schools that had already been selected for the purposes of piloting of instruments through the course of this project. We further excluded particularly small schools (fewer than 20 grade 1 enrolments) since many of these schools would practice multi-grade teaching rendering the scripted lesson plans less appropriate. We also excluded particularly large schools (more than 180 grade 1 enrolments) to limit intervention costs. Three more schools were excluded after the North West PED checked our list of schools and found specific problems with these schools (e.g. the school had been closed down, or a particular conflict around school management was occurring in a school). After all of these exclusions 235 eligible schools remained. Using a random number generator, we then excluded 5 schools, which we retained as possible replacement schools. Thus we obtained the sampling frame of 230 schools.

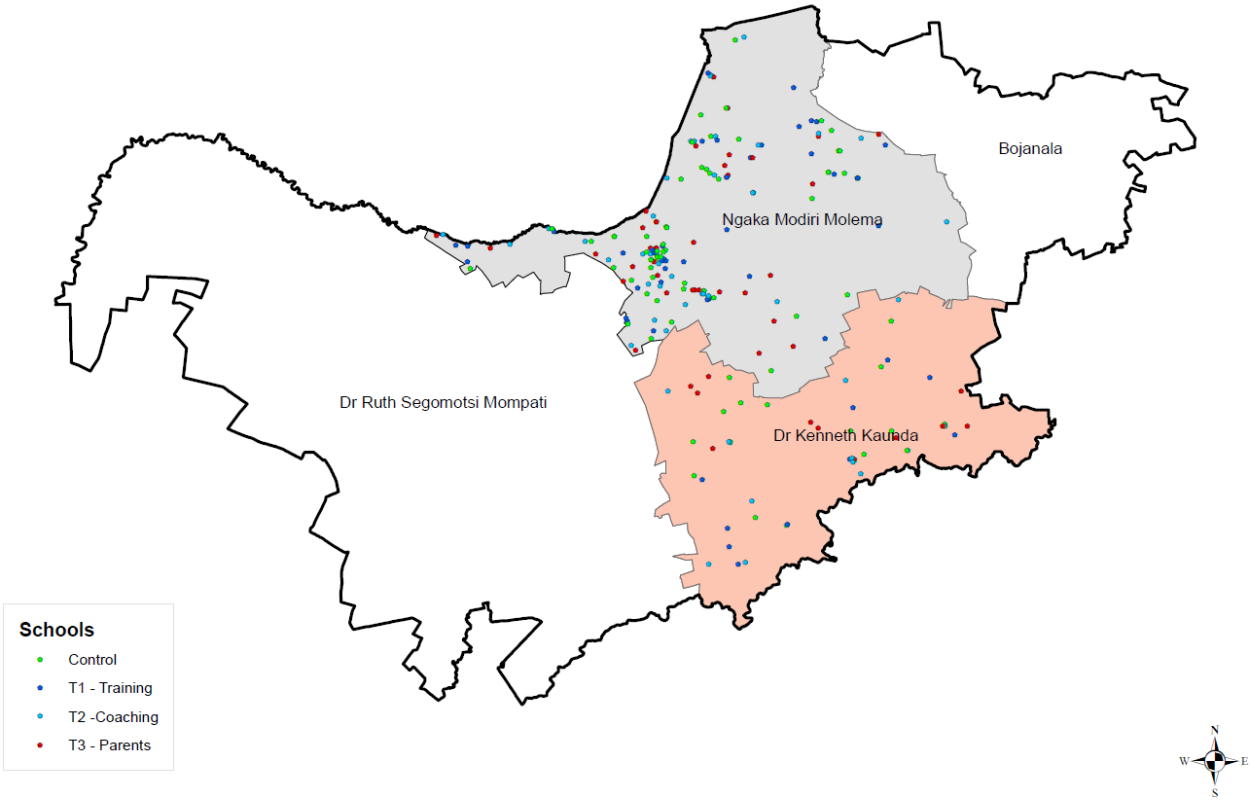
To increase power and assure balance between treatment arms, we performed stratified randomization. We created 10 strata of 23 similar schools based on school size, socio-economic status, and previous performance in the Annual National Assessments. Within each stratum, we then randomly assigned 5 schools to each treatment group and 8 to the control group. Thus we randomly assigned 50 schools to each treatment and 80 to the control. Given that we collect data on 20 grade 1 learners per school, this sample should be sufficient to identify a minimum effect size of 0.21 standard deviations when comparing a treatment group with the control group and a minimum effect size of 0.23 standard deviations when comparing two treatment groups. These calculations assume a 95% confidence interval, an alpha value of 0.8, an intra-class correlation coefficient (rho) of 0.3 and a correlation between pre- and post-test scores of 0.7. Figure 4 presents a schematic diagram to describe the sampling procedure that was followed.

Figure 4: Diagram showing sampling procedure



The following map shows the schools participating in the EGRS and indicates the treatment status of each school. Note that a few schools are not shown on the map due to missing or inaccurate GIS codes.

Figure 5: Map of North West province showing schools by treatment assignment



INSTRUMENT DEVELOPMENT AND PILOTING

The Research Team worked closely with the HSRC to develop four survey instruments for the baseline data collection: a learner test, a school principal questionnaire, a teacher questionnaire and a parent/guardian questionnaire. The learner test was designed in the spirit of the Early Grade Reading Assessment (EGRA) to be administered orally by a fieldworker to one child at a time. The test instrument used parts of the EGRA for Setswana, which had already been developed in South Africa. The letter recognition fluency, word recognition fluency and sentence reading components of the test were based on the Setswana EGRA instrument. A picture comprehension test (or expressive vocabulary test) was included since this was expected to be an easier pre-literacy skill testing vocabulary, and thus useful for avoiding a floor effect at the start of grade 1 when many children are not expected to read at all. A phonemic awareness test component was also added. Similarly, a digit span memory test was included – this involved repeating by memory first two numbers, then three, and so forth up to six numbers, and the

same 5 items for sequences of words. The logic of including this test of working memory is that it is known to be a strong predictor of learning to read. Thus, when estimating the impact of the three interventions after endline testing we can include as a control variable a measure of the child's working memory at baseline and in this way improve the precision of treatment effect estimates.

The school principal, teacher and parent questionnaires were designed in order to collect information to be used in the measurement of heterogeneous treatment effects (i.e. differential impact across relevant sub-groups of schools or learners) and to measure changes in intermediate outcomes along the hypothesized causal chain for each intervention. The parent questionnaire was sent home with those learners who were tested and then brought back to the school on a later day, to be collected by the fieldworkers on a return visit. In addition, a data linkage form was developed upon which all learner names and unique identifier numbers were linked to the appropriate teacher unique identifier and teacher name. All these instruments and the entire data collection process were piloted in 5 schools on the 3rd and 4th of September 2014. Following lessons learnt from the piloting, revisions were made to the instruments.

ETHICAL CLEARANCE

The methodology, with the intended instruments related to the baseline data collection, was formally submitted to the HSRC's Research Ethics Committee in February 2014. The project was approved in principle (i.e., provisionally) on 24 March 2014, pending submission of the final field-test and baseline instruments and site permissions. The relevant field-test documents were submitted and approved on 29 August 2014. Subsequently, after final revisions to the procedures and instruments for the baseline data collection, and submission of final site permissions along with an application for recertification for another year, ethics clearance was provided on 21 January 2015 for the baseline data collection.

DATA COLLECTION

Baseline data collection comprised visits to all 230 EGRS schools (150 treatment and 80 control schools) in order to assess the Setswana language proficiency of 20 Grade 1 learners per school (4 600 learners in total). The HSRC hired a fieldwork agency to recruit fieldworkers and manage their transport to schools during the fieldwork. A total of 60 fieldworkers were recruited, comprising 30 former teachers (to conduct the learner testing) and 30 other fieldworkers who were not necessarily education-specific. The plan was for fieldwork to be conducted in teams of 2, with one fieldworker conducting the learner testing and the other administering the school principal and teacher questionnaires, all in the course of a 1-day visit to each school. The HSRC were directly responsible for the printing and packaging of all instruments and passed these on to the fieldwork agency. The HSRC facilitated the training of the fieldworkers. This was initially a 1-day session. However, after some problems were evident on the first day of data collection a decision was taken to recall all fieldworkers for an additional 1-day re-training.

Data collection occurred between the 4th and the 24th of February. Monitoring of fieldwork occurred at two levels. Firstly, the HSRC sent monitors to observe fieldwork in a randomly selected (by the Research Team) 10% of schools, i.e. 23 schools. Secondly, the DBE made telephone calls to school principals to find out about how fieldwork had occurred at the school. Reports on both levels of monitoring were compiled.

A number of challenges were experienced during the data collection. Firstly, there were some problems with respect to the logistics of school visits. Although all schools should have known about their participation in the EGRS through a set of meetings with all principals at the end of 2014 and through letters from the NW PED, the telephone numbers for schools obtained through the DBE's EMIS data were in some cases incorrect or outdated. To add to this problem, the fieldwork schedule of which fieldworkers should attend which schools on which days, as arranged by the subcontracted fieldwork agency was regularly updated resulting in appointments with schools either not being set up or set up rather late. Fortunately, this did not lead to any outright refusals from schools to participating, and those few schools where initial refusal occurred were re-visited on a later day. Another challenge was that on some occasions fieldworker transport was not efficient so that a team of fieldworkers arrived late at school. This would have compromised the quality of data collection at such a school due to time constraints. A further challenge experienced is of incomplete return of instruments by fieldworkers, possibly partly due to late arrival at schools.

The intention was for the fieldworkers to randomly sample 20 learners per school, using a specified procedure. The fieldworker was to obtain from the teachers the full list of children enrolled in grade 1, putting one class list below the next if a single grade list was not provided. The fieldworker was to tally the total number of children and divide this number by 20. The answer was then to be rounded up to the nearest whole number, n. The fieldworker was then to start with the third learner and select every n^{th} learner for inclusion in the sample. Upon reaching the end of the list the fieldworker was to go back to the top of the list and continue selecting every n^{th} learner, not counting previously selected learners, until 20 learners have been selected. Monitoring of fieldwork indicated that in a few cases, the fieldworkers may not have followed the procedure perfectly. However, there was no evidence of systemic sampling of learners through anybody's recommendation. If a fieldworker attempted the procedure but misunderstood it the resulting sample should still be effectively random. Therefore, there is no reason to expect a systematically stronger or weaker sample to have been selected, and there is certainly no reason to expect any differences in sampling across treatment groups.

The following tables provide a sense of the data completeness as far as instrument returns is concerned. Learner testing occurred in all 230 schools, providing a realized sample of 4539 learners. Table 4 shows that in the majority of schools (204 out of 230 schools) exactly 20 learners were tested and the data successfully captured. In 4 schools there were 21 learners tested. It is not clear why this occurred. It may have been a counting error by the fieldworker or perhaps a small school only had 21 learners and it was felt that a single learner should not be left out. The few cases of 15, 16, 17 and 19 learners tested is not unexpected since there are

known to be some small schools in the sample. Schools with fewer than 20 grade 1 enrolments in 2014 were excluded from the sampling frame; but we know from administering the interventions that some of the schools have lower enrolments in 2015. Although not impossible, it does seem unlikely that schools would only have had 9 or 10 grade 1 enrolments. To some extent, therefore, incomplete fieldwork may have led to fewer than 20 learners were tested.

Table 4: Number of learners successfully tested per school

Number of learners	No of schools with this number of learners
9	1
10	1
15	2
16	4
17	2
19	12
20	204
21	4
Total	230

As Table 5 indicates, the return of parent questionnaires was rather erratic. The parent questionnaire was sent home with tested children and was meant to be brought back to the school and then collected on a later day by the fieldwork agency. The weakness of this method is that children may not always bring the questionnaire back. However, it is more reliable than asking children themselves about hoe characteristics. It is concerning, however, that no parent questionnaires were returned in 49 schools. This is most likely a reflection of poor fieldwork or of a lack of cooperation from school staff. Importantly, there was no significant pattern of instrument return across treatment group, not that one would expect that given that fieldworkers were blind to treatment allocation and that interventions had not yet commenced. For those schools where parent questionnaires were returned the return rates were not too bad, as described in Table 6. About 60% of schools had return rates of greater than 50% (i.e. 10 learners or more). If one excludes, the schools where no parent questionnaires were returned (not shown in Table 6), then about 75% of schools had return rates of 50% or more, and about 60% of schools had return rates of at least 75% (i.e. 15 learners).

Table 5: Numbers of returned learner tests and parent questionnaires

	Learner tests		Parent Questionnaires	
	No Students	No schools	No Students	No schools
Control	1575	80	856	62
Treatment 1	983	50	559	42
Treatment 2	982	50	569	41
Treatment 3	999	50	500	36
Total	4539	230	2484	181

Table 6: Number of parent questionnaires returned per school

Number of parent questionnaires	No of schools with this number of learners	Cumulative percentage
0	49	21.3
2	2	22.17
3	1	22.61
4	7	25.65
5	4	27.39
6	3	28.7
7	3	30
8	7	33.04
9	9	36.96
10	9	40.87
11	10	45.22
12	9	49.13
13	11	53.91
14	15	60.43
15	18	68.26
16	15	74.78
17	8	78.26
18	18	86.09
19	18	93.91
20	14	100

Two separate instruments were supposed to be administered to all teachers in grade 1. The first instrument is the teacher questionnaire, which collected a variety of information about teacher demographics, attitudes and practices. The second instrument was a short reading fluency test for teachers (to be described in more detail in a later section of this report). A questionnaire was also given to school principals to complete. Table 7 shows the numbers of teacher and principal instruments returned. The principal return rate is straightforward since one expects one questionnaire per school. In 14 schools no principal questionnaire was completed and returned

by the fieldwork agency. There were 326 teacher questionnaires returned that could be linked to learners.[1] There were also 320 teacher fluency tests returned. However, in some of these cases the teacher questionnaire data could not be linked to the teacher fluency test data. This may partly reflect inaccurate personal details and incorrect application of unique identifiers by the fieldworkers. However, manual investigation of these unmatchable cases would suggest that there may have been some teachers who only completed one of the instruments. 286 teachers were successfully matched across the two instruments. As Table 7 indicates, although over 300 teacher questionnaires and fluency tests were returned, since more than one teacher could be interviewed per school, the number of schools in which at least one teacher was surveyed was unfortunately less than the intended 230 schools. In only 198 schools was at least one teacher questionnaire returned. The fluency test was successfully administered and captured in 194 schools. It is possible that teacher refusal to be tested could have contributed somewhat to the non-return of teacher fluency data. However, the appropriate procedure for the fieldworker to follow in the case of refusal to participate was to return the test instrument with a field indicating whether the teacher was willing to participate – 40 teachers were not willing according to this variable and thus had missing data on the test score variables.

Table 7: Numbers of returned teacher and principal questionnaires

	Teacher Questionnaire		Teacher Fluency test		Principal Questionnaire
	No Teachers	No schools	No Teachers	No schools	No schools
Control	112	70	107	65	71
Treatment 1	72	43	73	44	49
Treatment 2	77	46	80	45	48
Treatment 3	65	39	60	40	48
Total	326	198	320	194	216

In summary, it would appear that imperfect fieldwork contributed to a lower than intended return rate of survey instruments. Fortunately, non-return is not systematically related to treatment assignment. It is also fortunate that the main priority of learner testing was generally fairly complete. Rather a lot of non-return occurred for the parent, teacher and principal instruments. Moreover, even when instruments were returned there was rather a lot of item non-response. This will limit the evaluation analysis once midline and endline data are collected in several ways. Firstly, the main impact estimation model will not include many parent, teacher and school covariates as controls. The value of such controls is to slightly improve statistical power when estimating the treatment effects. However, this power gain is rather marginal so the loss is not too bad. Moreover, in an RCT setting where the source of variation in treatment assignment is strictly exogenous by design one would expect no bias to have to control for through the

[1] In fact a few more teacher instruments were returned and captured but due to incomplete identification information these could not be linked to learners at schools and were therefore excluded from the merged dataset and this analysis.

inclusion of covariates. Therefore, the inclusion of many covariates in an RCT regression model is in any case not always favoured by analysts. A more worrying limitation is that missing information on baseline characteristics will mean a reduced effective sample size when estimating certain heterogeneous treatment effects and when estimating impacts on intermediate outcomes, such as teacher attitudes and practices. One way to mitigate these problems will be through collecting much of the same information in the midline and endline surveys (November 2015 and November 2016). Certain information, such as teacher age, is not expected to change in response to treatment and can therefore be used in the estimation of heterogeneous treatment effects even if the information was collected after interventions commenced. Treatment effects on intermediate outcomes can be estimated without controlling for baseline characteristics since there is no reason to expect any differences between treatment groups other than because of the causal effect of the interventions. The disadvantage is that power is reduced through the lack of controlling for baseline variation.

A number of steps are being taken in the midline data collection (scheduled for 26 October – 13 November 2015) to improve the data collection. The Terms of Reference for the subcontracting of a fieldwork agency is now much more detailed with respect to fieldworker selection criteria, conditions around approval of and payment for deliverables, and overall functionality criteria for the fieldwork organization. The entire procurement process of the fieldwork agency for midline data collection is happening in good time to ensure adequate lead up time to the data collection. Instead of a single day of fieldworker training, there will be a three-day training programme for fieldworkers including a practice round of data collection (with monitoring and feedback) at five schools not included in the project. The Terms of Reference has specified that exactly 40 fieldworkers should be recruited, 20 of whom will administer the learner tests and must have expertise in early grade teaching. The fieldwork schedule also needs to be submitted well in advance to the HSRC with schools already having been contacted and appointments fixed for specific days made. This process of communicating with schools is also likely to be smoother since we now have an updated database of contact information, which the DBE compiled using information collected in baseline questionnaires and by the implementing agent for interventions. Finally, extensive revisions have been made to the midline instruments, especially the shortening of the school principal and teacher questionnaires.

DATA CAPTURING AND CLEANING

Questionnaires were unpacked and data was cleaned within the HSRC by in-house data capturers. Six separate datasets were thus captured, corresponding to the different instruments. The six datasets were the data linkage file (linking learner, teacher and school unique identifiers), the learner test data file, the parent questionnaire data file, the school principal data file, the teacher questionnaire data file and the teacher reading fluency test data file. A preliminary version of these datasets was provided by the HSRC to the Research Team. Initial analysis of this data identified several data issues. These included one school that was missing from the data, some obvious mistakes in unique identifiers of learners and teachers, one data file that had mixed up the school identifier numbers, etc. The Research Team then sent a set of queries to the HSRC, who in turn investigated these issues. After some re-capturing and

cleaning, the HSRC then provided the final baseline datasets to the Research Team. Some of the data queries had been satisfactorily resolved (e.g. the "missing" school was found – it had initially been confused by the data capturers with another school with a very similar name), while other issues could not be fully resolved as they stemmed from fieldworker errors in capturing information.

Even after receiving the final datasets from the HSRC, the Research Team needed to do additional data cleaning, which was clearly needed once attempting to merge the various datasets. For example, there were a few duplicate learner IDs that needed to be adjusted by manually looking at learner names and surnames and comparing with the linkage data file. Similarly, a number of EMIS numbers (official school unique identifier) were incorrect in the parent questionnaire dataset. These were easily identified and corrected. The data cleaning done by the Research Team is recorded in STATA do-files, which will be made publicly available at the end of the project.

BASELINE RESULTS

LEARNER TEST SCORES

The baseline learner test instrument, which will be made publicly available at the end of the project, was adapted from the Setswana Early Grade Reading Assessment (EGRA). The tests were therefore administered to one child at a time. The average time per test was about 15 minutes. In view of the fact that the baseline assessment took place at the very start of grade 1, one modification was the inclusion of some items which could be described as assessing pre-literacy skills so as to be sure to avoid a floor effect (where a substantial proportion of learners score zero or close to zero on the test as a whole). Section A thus consisted of 10 picture comprehension items, which test expressive vocabulary – a skill which should be fairly well developed by the start of primary school. Six of these items were pictures of well-known objects, such as a car and a spoon. The remaining four pictures displayed some sort of action, such as a bird flying or a child sleeping. In each case, the learner was asked to say the Setswana word for the object or action.

One problem encountered in the scoring for Section A (as for Sections C and D), was that the fieldworker was supposed to mark each of the ten items correct or incorrect and also to indicate the total score out of ten. However, in some cases the sum of the item scores did not tally to the total score recorded by the fieldworker. This occurred for 200 learners out of the total of 4540 learners. In cases where the fieldworker left all individual items blank but entered a valid total score, we used that total score. In cases where the total score was missing we imputed the sum of the individual item scores as the total score. In cases where the difference between the calculated sum of individual scores was 1 or 2 points away from the recorded total score we decided to use the calculated sum of scores under the assumption that this was probably a fieldworker counting error. In cases where the difference was greater than 2 points it is unlikely

that this could be a counting error and therefore we used the recorded total score under the assumption that scoring the individual items was erratically done.

The summary statistics for all items in sub-tests A and C are presented in Table 8, and for sub-test D in Table 9. The summary statistics for the total scores per subtask as well as an overall composite test score are shown in Table 10. For Section A it can be seen that most learners did rather well in this section. Items 4 and 9 were the hardest items in Section A with 61% and 62% of learners getting the answers correct, respectively. The average score out of 10 was 8.58. The inclusion of these easier items was deliberate since it was expected that the majority of learners would struggle with the traditional EGRA items, which require some reading ability. Figure 6 confirms that the majority of students achieved scores of 8, 9 and 10 out of 10 and that there was a ceiling effect on this subtask. The figure also shows the distributions of scores for each of the four treatment arms. The distributions are virtually identical for each treatment group. This confirms the success of the randomization to ensure a well-balanced treatment assignment. One concern with Section A is the low Cronbach's alpha (0.52) that was obtained, indicating that the items are not combining to present a very reliable measure of an underlying construct. There were no individual items that were so problematic that if removed would increase Cronbach's alpha. This analysis of Cronbach's alpha is presented in Table 11.

Table 8: Summary statistics for items in sub-tests A and C

	Observations	Mean	Std. Dev.	Min.	Max.
Score for Item A1	4509	0.92	0.28	0	1
Score for Item A2	4487	0.96	0.19	0	1
Score for Item A3	4503	0.99	0.10	0	1
Score for Item A4	4384	0.61	0.49	0	1
Score for Item A5	4495	0.97	0.17	0	1
Score for Item A6	4441	0.85	0.36	0	1
Score for Item A7	4459	0.84	0.36	0	1
Score for Item A8	4493	0.96	0.20	0	1
Score for Item A9	4391	0.62	0.49	0	1
Score for Item CW1	4404	0.89	0.31	0	1
Score for Item CW2	4404	0.82	0.38	0	1
Score for Item CW3	4296	0.52	0.50	0	1
Score for Item CW4	4216	0.20	0.40	0	1
Score for Item CW5	4174	0.12	0.33	0	1
Score for Item CN1	4399	0.93	0.26	0	1
Score for Item CN2	4383	0.85	0.36	0	1
Score for Item CN3	4276	0.48	0.50	0	1
Score for Item CN4	4206	0.24	0.43	0	1
Score for Item CN5	4132	0.14	0.34	0	1

Table 9: Summary statistics for items in sub-test D

	Observations	Mean	Std. Dev.	Min.	Max.
Score for Item D1_1	4163	0.36	0.48	0	1
Score for Item D1_2	4131	0.26	0.44	0	1
Score for Item D1_3	4112	0.21	0.41	0	1
Score for Item D1_4	4094	0.19	0.40	0	1
Score for Item D1_5	4090	0.16	0.37	0	1
Score for Item D1_6	4070	0.13	0.34	0	1
Score for Item D1_7	4144	0.25	0.43	0	1
Score for Item D1_8	4158	0.28	0.45	0	1
Score for Item D1_9	4119	0.22	0.42	0	1
Score for Item D1_10	3999	0.13	0.33	0	1
Score for Item D1_11	3990	0.10	0.29	0	1
Score for Item D1_12	3978	0.08	0.28	0	1

Table 10: Summary statistics – Aggregate test scores

	Observations	Mean	Std. Dev.	Min.	Max.
Picture comprehension /10	4539	8.58	1.40	0	10
letters correct	4452	5.08	9.86	0	99
digit span words /5	4539	2.45	1.31	0	5
digit span numbers /5	4539	2.54	1.34	0	5
digit span total /10	4539	4.99	2.43	0	10
phonemic awareness /12	4539	2.17	3.11	0	12
words correct	4447	1.91	5.28	0	50
sentence reading comprehension /3	4539	0.73	1.24	0	3
number of sentence words correct /15	4539	1.22	3.38	0	15
combined score mean 0 SD 1 pca	4385	0.00	1.00	-2	5

Figure 6: Kernel Density Curves for Section A (expressive vocabulary) by treatment arm

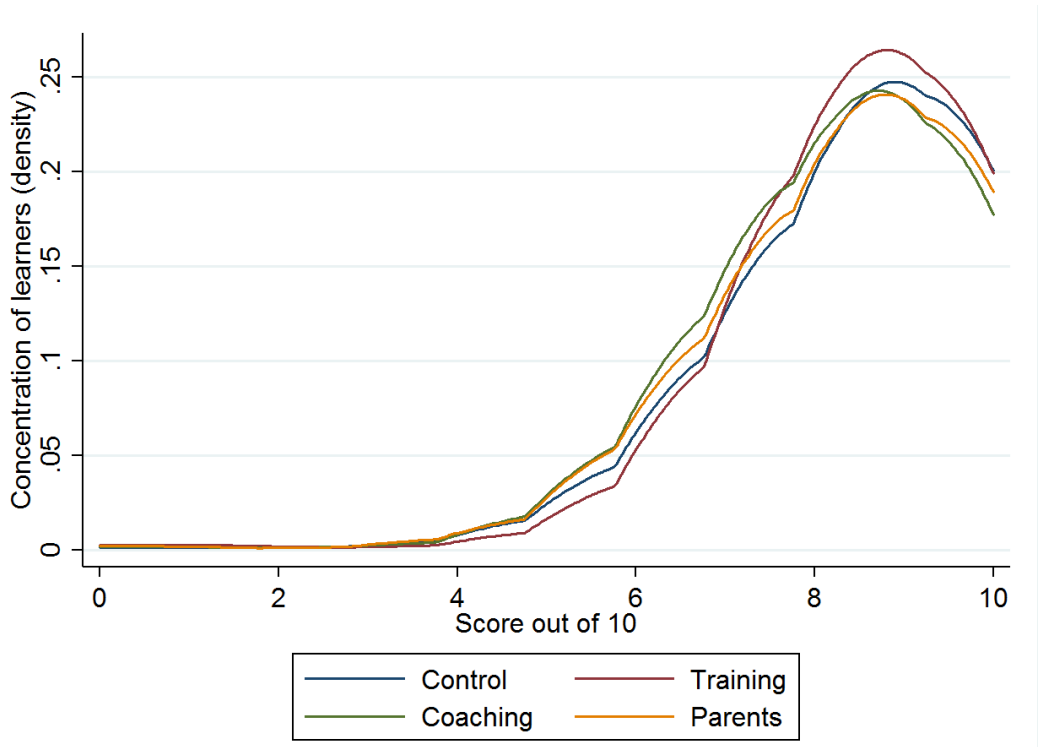
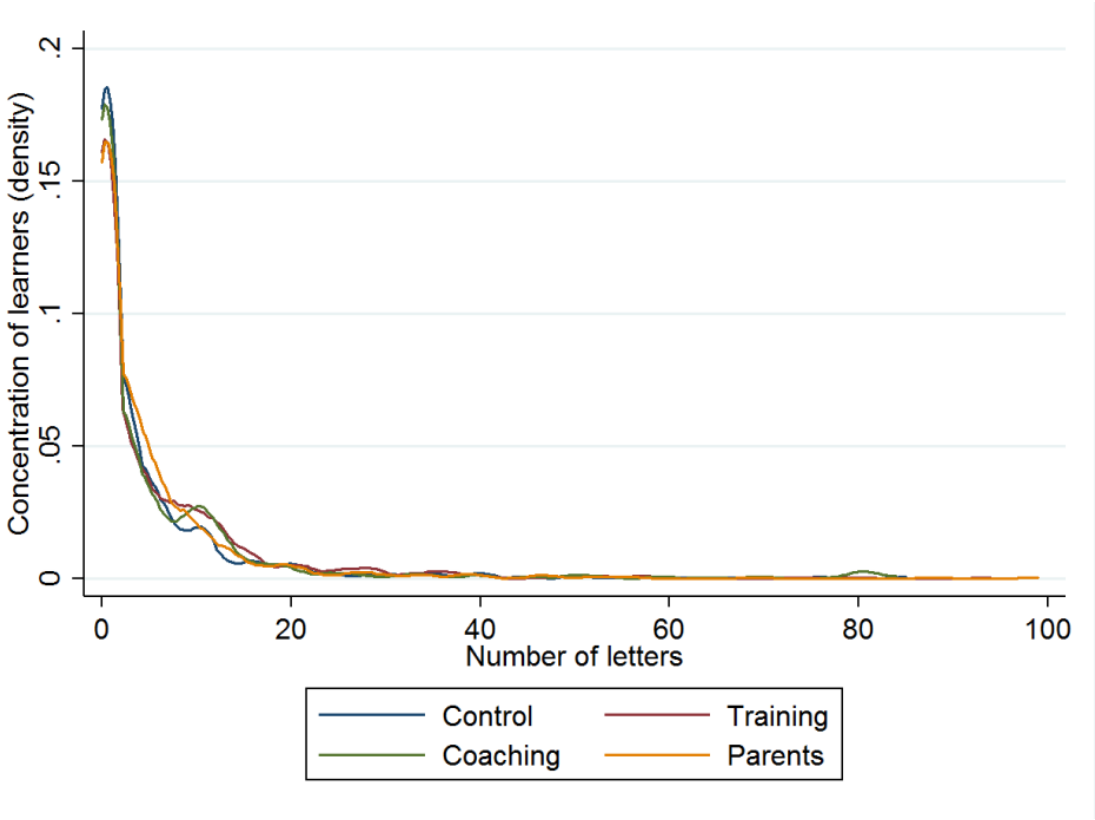


Table 11: Cronbach's alpha for Section A (expressive vocabulary)

Item	Observations	Sign	Item-test correlation	item-rest correlation	Alpha
A1	4509	+	0.43	0.25	0.49
A2	4487	+	0.33	0.19	0.50
A3	4503	+	0.31	0.24	0.51
A4	4384	+	0.54	0.21	0.51
A5	4495	+	0.34	0.22	0.50
A6	4441	+	0.43	0.18	0.51
A7	4459	+	0.51	0.27	0.48
A8	4493	+	0.40	0.26	0.49
A9	4391	+	0.63	0.34	0.45
A10	4488	+	0.43	0.29	0.48
Test					0.52

Section B was the letter recognition test, which is a conventional EGRA task. Learners were given 60 seconds to read as many letter sounds as possible. At the end of the 60 seconds the fieldworker captures the number of letters reached as well as the number of letters correct. As Table 10 indicates, the number of letters reached was typically far higher (averaging 29) than the number of correct letter sounds read (averaging 5). This is to be expected since the fieldworker moves the pointer along to the next letter if the learner has not provided an answer after three seconds. There was quite a substantial floor effect on this subtask, as indicated by Figure 7. About 42% of learners could not read and pronounce any letter sounds. As with Section A, the kernel density curves were virtually identical across the treatment arms. This confirms that the groups are well balanced on baseline.

Figure 7: Kernel Density Curves for Section B (letters correct) by treatment arm



Section C of the test was a digit span memory test designed to provide a measure of a child's working memory, which is known to be a strong predictor of learning to read. The item involved the fieldworker saying two unrelated Setswana words and the learner needed to repeat them back to the fieldworker from memory. The second item involved three unrelated words, the third had four words, the fourth had five words and item five involved six words. The next 5 items followed the same pattern but using numbers (spoken in Setswana). The rationale for including this sub-task is not because it is a reading outcome but because it is predictive of learning to read. This will be important for our final impact evaluation analysis where including baseline measures of cognitive ability, if these are well correlated with reading outcomes at the endline, can be expected to account for some of the variation in reading outcomes and thus increase the

precision with which we can measure the impact of the interventions. There was a good spread of achievement on these items. For the word span test, 89% of learners could successfully repeat the two-word sequence, with smaller proportions being able to repeat more words, down to only 12% who could repeat six words. With the numbers section, 93% of learners could successfully repeat the two-number sequence while only 14% could repeat the six digit sequence. The reliability of Section C is somewhat better than that observed for Section A, as the analysis of Cronbach's alpha in Tables 12 and 13 demonstrate. Figure 8 demonstrates that neither a floor effect nor a ceiling effect exists for Section C. Rather, the distribution of scores approximates a normal distribution, which is encouraging for the purposes of providing a good baseline measure of learner cognitive ability. The figure also confirms good balance across treatment groups.

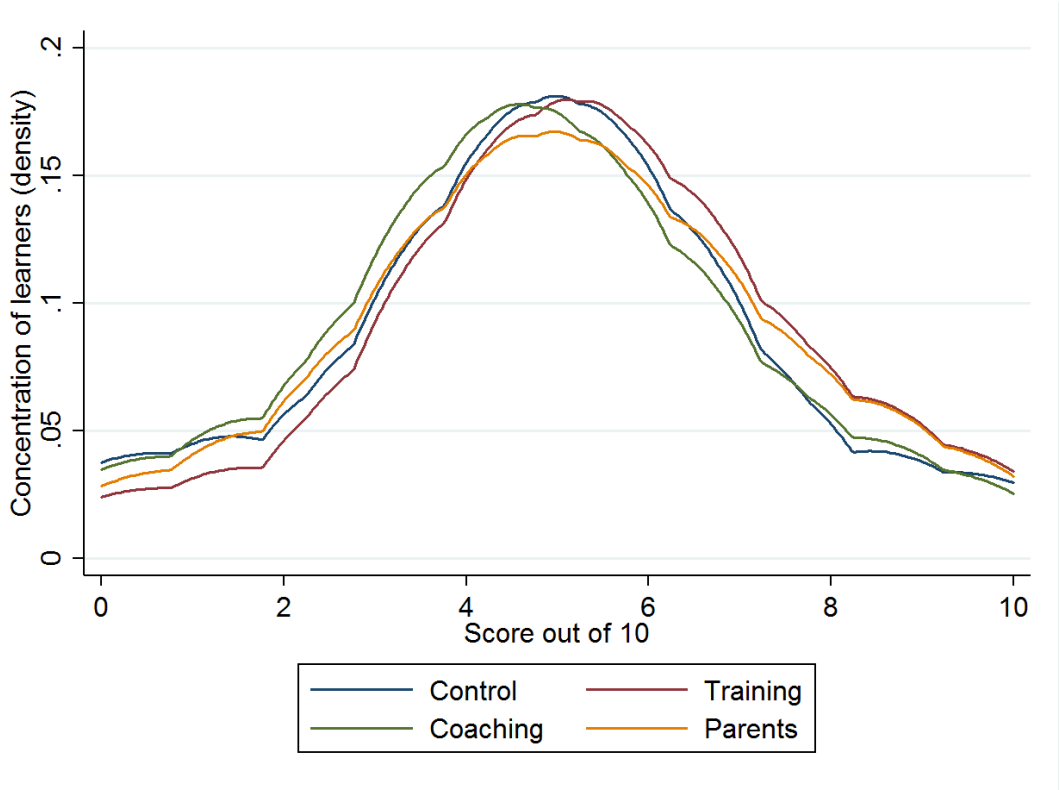
Table 12: Cronbach's alpha for Section C.1 (short-term memory – words)

Item	Observations	Sign	Item-test correlation	item-rest correlation	Alpha
CW1	4404	+	0.58	0.41	0.66
CW2	4404	+	0.66	0.46	0.64
CW3	4296	+	0.78	0.55	0.61
CW4	4216	+	0.70	0.48	0.64
CW5	4174	+	0.59	0.40	0.67
Test					0.70

Table 13: Cronbach's alpha for Section C.2 (short-term memory – numbers)

Item	Observations	Sign	Item-test correlation	item-rest correlation	Alpha
CN1	4399	+	0.52	0.36	0.70
CN2	4383	+	0.63	0.43	0.68
CN3	4276	+	0.79	0.56	0.64
CN4	4206	+	0.76	0.56	0.63
CN5	4132	+	0.67	0.49	0.66
Test					0.71

Figure 8: Kernel Density Curves for Section C (short-term memory) by treatment arm



Section D tested phonological awareness, in three different ways. For the first six items the fieldworker read a Setswana word (e.g. “pitsa”) out loud and the learner had to break the word down into its smallest sound components or phonemes. For items D7 to D9 the fieldworker read a word out loud (e.g. “sega”) and the learner then had to suggest another word beginning with the same two letter sounds (e.g. “seba”). For items D10 to D12 the fieldworker read a word out loud (e.g. “yona”) and the learner then had to suggest another word ending with the same two letter sounds (e.g. “bana”). Table 9 shows that most learners struggled with this subtask, especially with items D10 to D12. The average score out of 12 on Section D was 2.17. There was also a floor effect, as can be seen in Figure 9. Again, the scores appear balanced across treatment groups. Members of the Research Team have observed this subtask being administered and have been somewhat concerned because it is rather difficult for fieldworkers to implement and confusing for learners to understand. While it is designed to test phonological awareness, which is an important component in learning to read, the “rules of the game” are difficult for children to grasp. For example, sometimes children break the word down into syllables rather than the smallest sound components. In such cases the child’s actual phonological awareness may be underestimated due to not understanding what is being requested of them. Despite these concerns the test produced a high degree of reliability as measured by Cronbach’s alpha, which was 0.90 (Table 14).

Figure 9: Kernel Density Curves for Section D (Phonological Awareness) by treatment arm

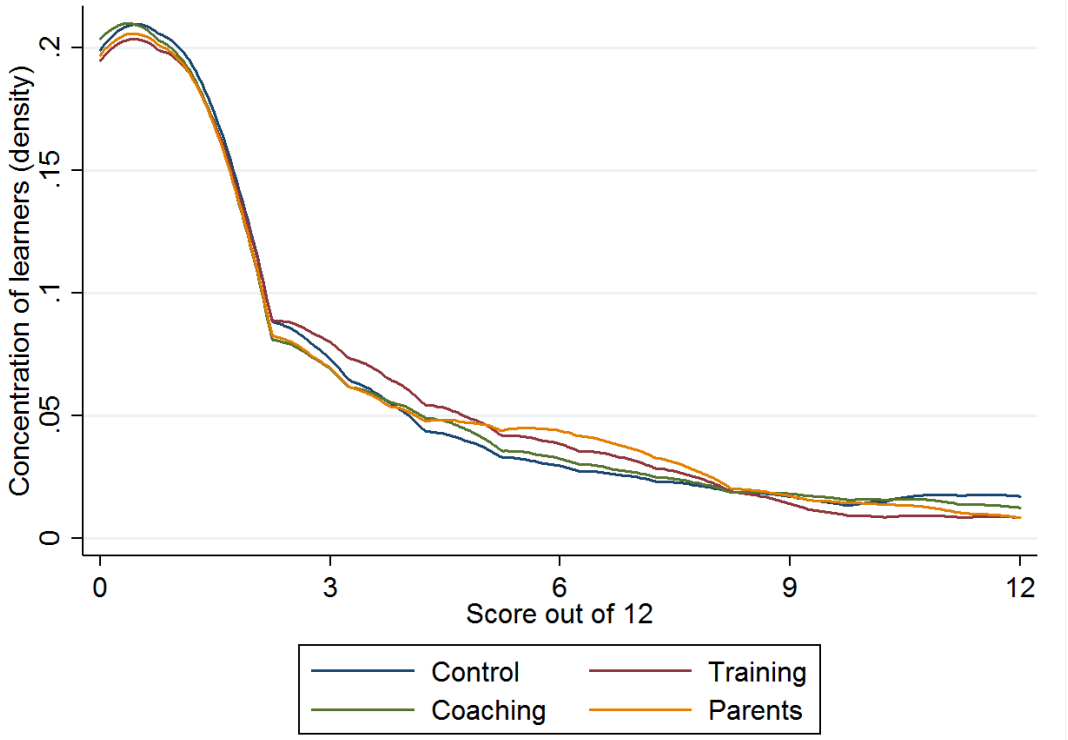
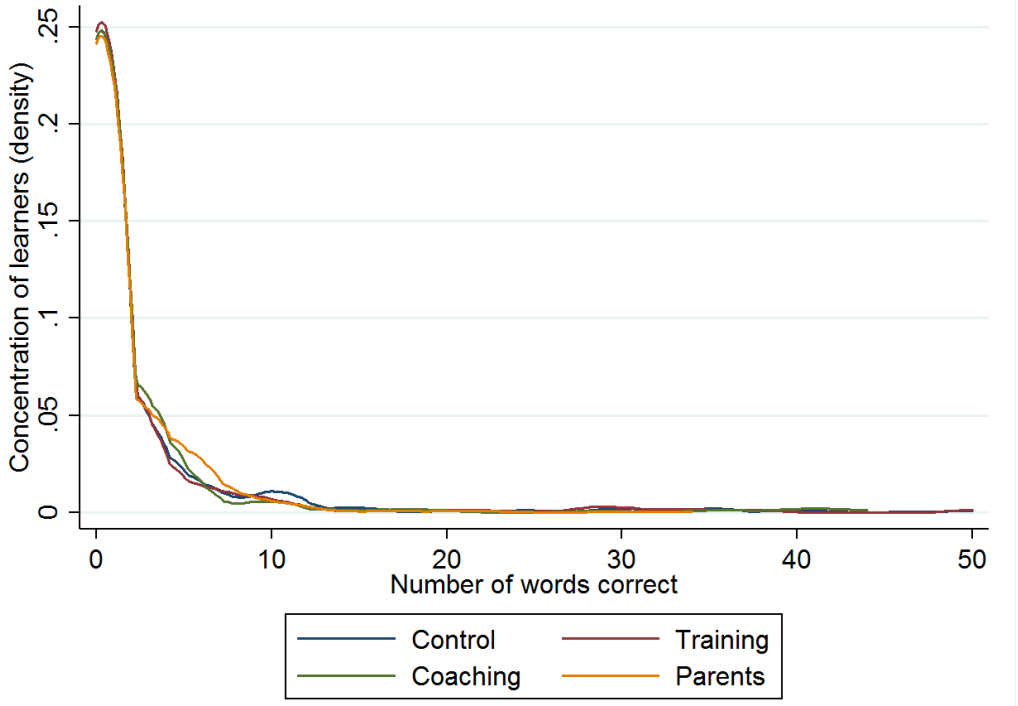


Table 14: Cronbach's alpha for Section D (Phonological Awareness)

Item	Observations	Sign	Item-test correlation	item-rest correlation	Alpha
D1_1	4163	+	0.71	0.62	0.90
D1_2	4131	+	0.79	0.74	0.89
D1_3	4112	+	0.82	0.77	0.89
D1_4	4094	+	0.78	0.73	0.89
D1_5	4090	+	0.79	0.74	0.89
D1_6	4070	+	0.74	0.69	0.89
D2_7	4144	+	0.70	0.62	0.90
D2_8	4158	+	0.65	0.56	0.90
D2_9	4119	+	0.64	0.55	0.90
D3_10	3999	+	0.67	0.60	0.90
D3_11	3990	+	0.69	0.63	0.90
D3_12	3978	+	0.67	0.61	0.90
Test					0.90

Section E was the word recognition test. As with Section B, the learner was given 60 seconds to read as many words as possible out of a maximum of 50 words. As would be expected at the start of grade 1, performance was low on this subtask and there was a substantial floor effect, as seen in Figure 10. The average score on this subtask was 1.91 correct words read. Balance was again good.

Figure 10: Kernel Density Curves for Section E (word recognition) by treatment arm



Section F consisted of three short sentences to be read by the learner. The learner was awarded a mark for every word that was correctly read. Altogether, there were 15 words across the three sentences. As Figure 11 indicates, approximately 80% of children were not able to read any of the words. A small proportion of children (about 4%) were able to read all 15 words. As before, balance was good. After reading each sentence, the learner was asked a comprehension question about that sentence. All answers were one-word answers. The average score out of 3 on the comprehension questions was 0.73 with about 73% of learners scoring zero. Interestingly, 21% of learners scored 3 out of 3 with very few learners scoring 1 or 2 out of 3. It would appear that learners can either read a sentence with comprehension or not and that including all three items did not add much value over and above the first item. For this reason, the midline test instruments will have only two sentences with the second being more complex than the first. Figure 13 indicates the positive association between word recognition and comprehension, as is expected. Similarly, a positive correlation was observed between letter recognition (section B) and comprehension and between word recognition (section E) and comprehension.

Figure 11: Number of words correct in Section F (sentence reading)

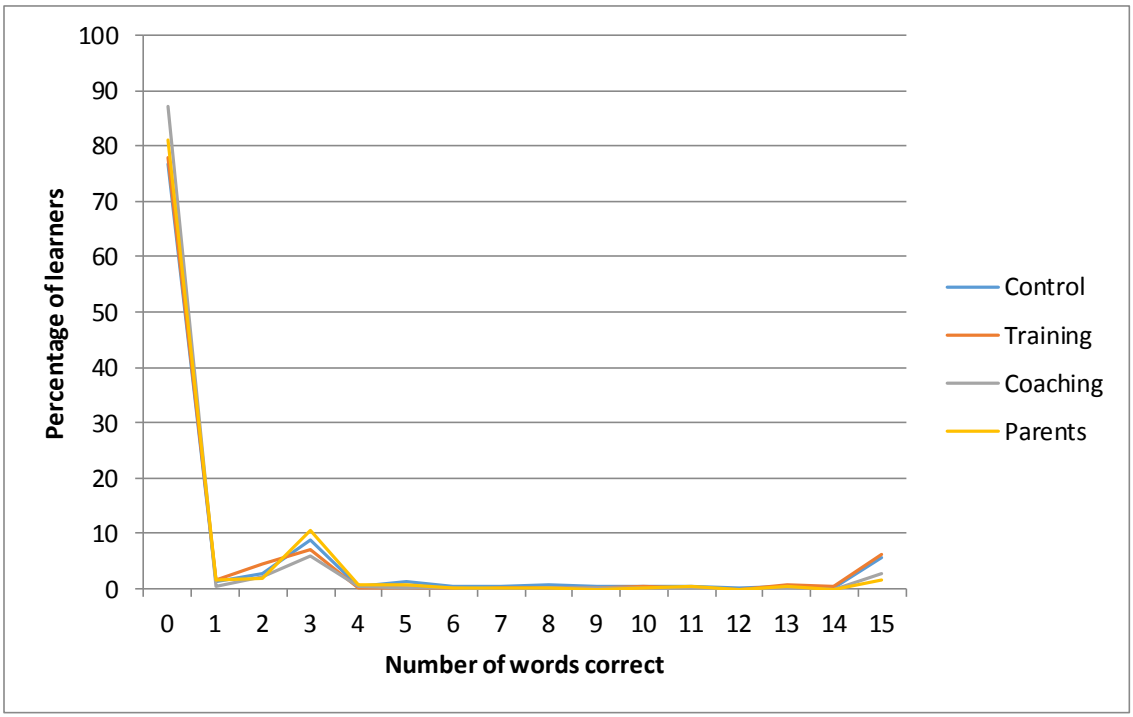


Figure 12: Percentage of learners scoring 0, 1, 2 and 3 for Section F comprehension

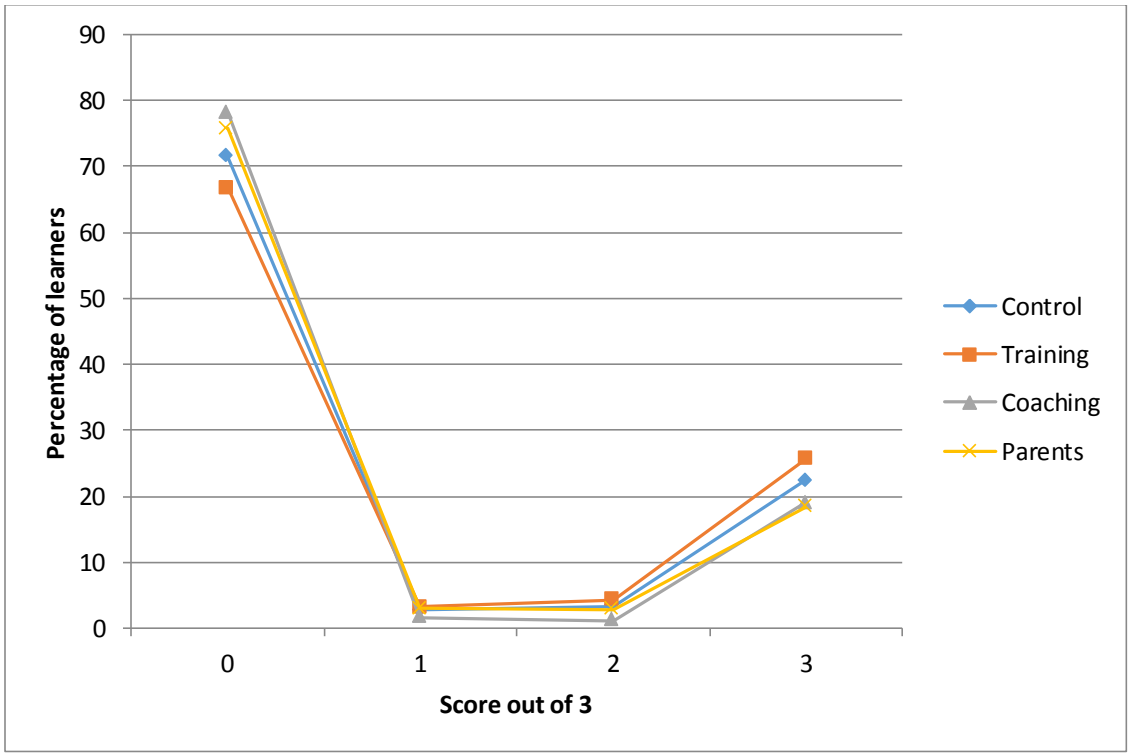
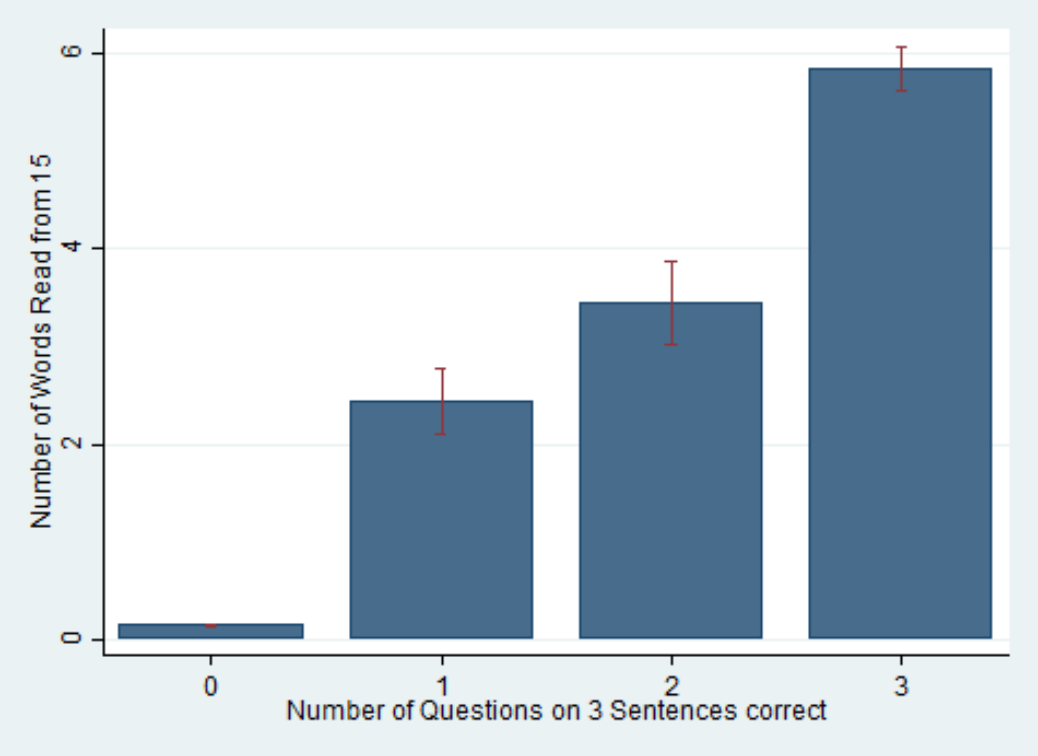


Figure 13: Relationship between words read and comprehension



In order to gain a sense of which sections of the test provide the best information we fitted a 1-parameter Item Response Theory (IRT) model treating each subtask as an individual item. Figure 14 shows item information functions for each subtask. Section A, where a ceiling effect was observed, provides some weak information about the lower part of the ability distribution and little information to distinguish amongst higher ability students. In contrast, Section E provides a lot of information at the high end of the ability distribution but little information about weaker learners. Sections C, D and F provide good information about the upper middle parts of the distribution but little information about the very bottom or very top of the distribution. Figure 15 aggregates all this into a single test information function. This confirms that the test does provide some information to distinguish between students at all parts of the ability distribution, but the information is best amongst the upper middle part of the distribution.

Figure 14: Item Information Functions from a 1-parameter IRT model

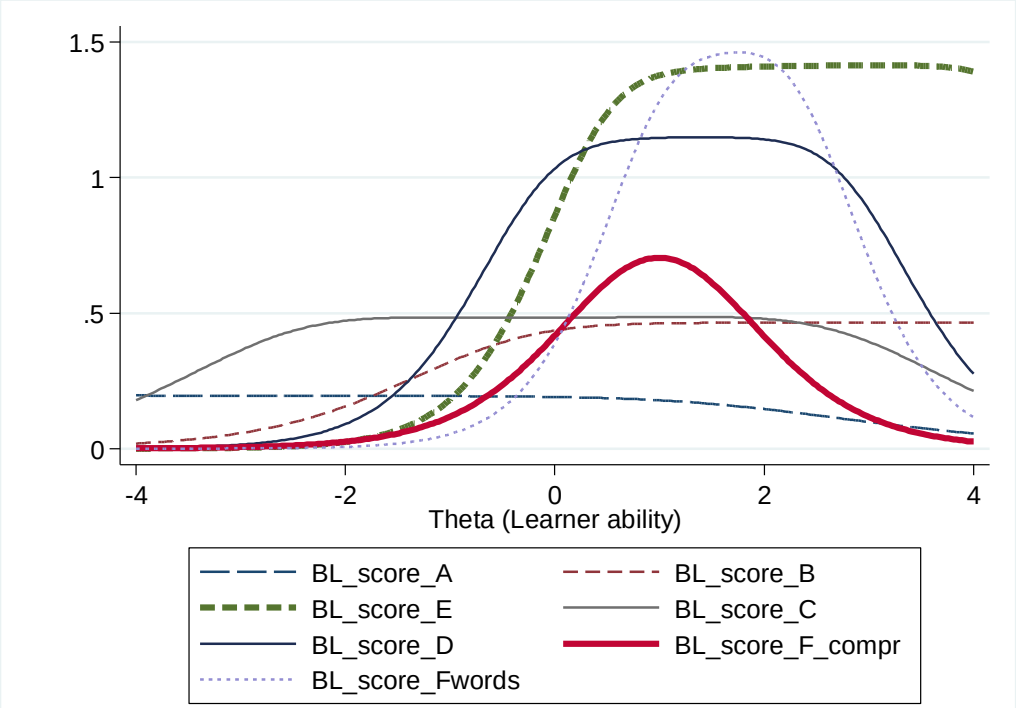
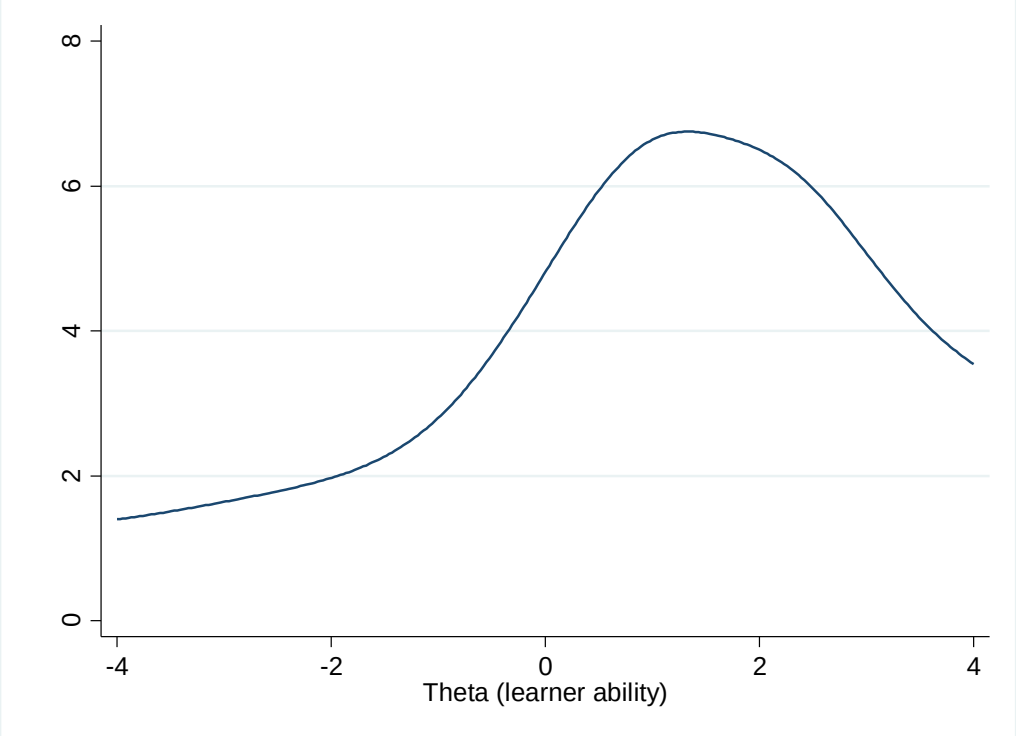
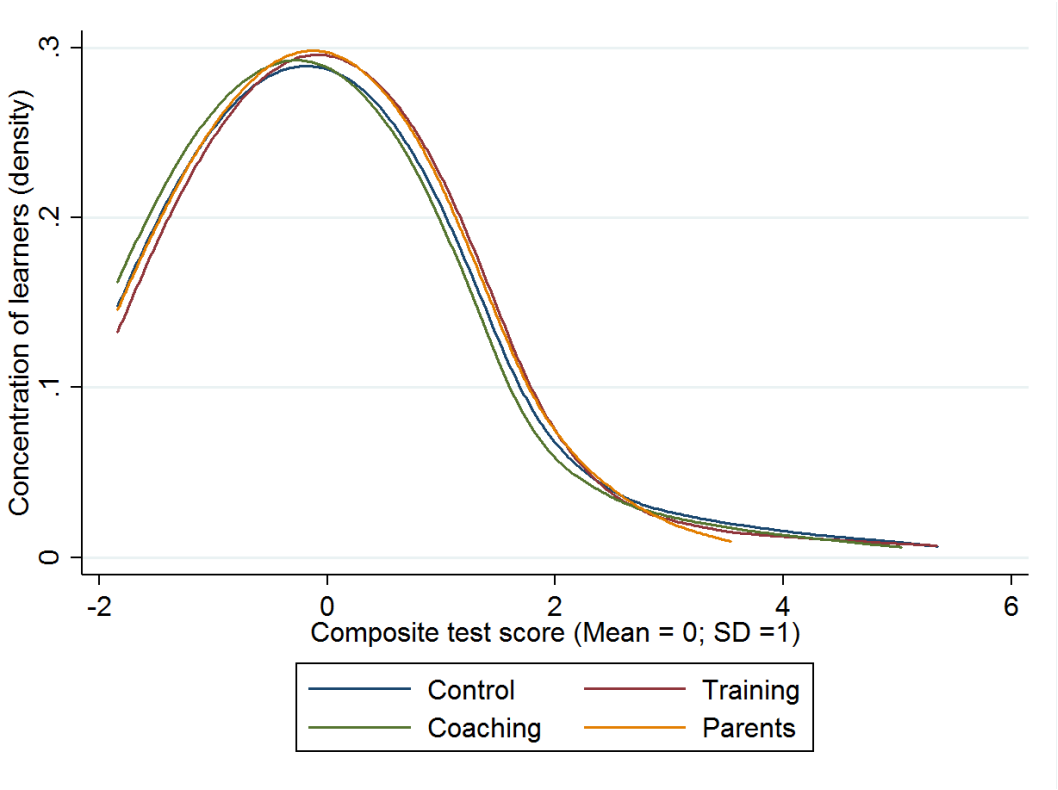


Figure 15: Test Information Function from a 1-parameter IRT model



Throughout the various subtasks there has been some evidence of "floor" and "ceiling" effects for particular subtasks. This is an important consideration, because we want to measure a change in learning outcomes across the whole distribution of pupils. Our statistical power is reduced if there is limited variation in baseline. For example, there is a "ceiling effect" in the vocabulary test where a large proportion got all the answers correct. Similarly, there is a "floor effect" on the number of letters that a pupil correctly read – the majority of the pupils did not get a single letter correct. Nonetheless, when we combine all the different learning measures into one composite score, using principal component analysis, we find a good normal distribution of learning outcomes (shown in Figure 16).[2] This is encouraging, because it means we will be able to detect a change in learning outcomes for *all* pupils across the distribution, and not only the best or worse-performers.

Figure 16: Kernel Density Curves for composite test score by treatment arm



[2] In calculating a composite score one needs to decide how much weight to attach to each subtask in the test. One cannot calculate simply add each subtask's score together, since one subtask may have had more items but should not necessarily carry more significance than another subtask. Therefore, we ran Principal Components Analysis (PCA) on the subtotals for each subtask, treating Section F comprehension as a separate score from Section 5 words correct. In PCA the variation within all variables included is analysed and those linear combinations capturing the most common variation amongst variables are identified. It is assumed that the linear combination, referred to as a principal component, which captures the most common variation amongst the variables included represents the underlying construct of interest. In this case we might think of the primary underlying construct being measured as reading ability or pre-reading ability. The weight given to each variable when calculating the total composite score is then determined by the extent of that variable's correlation with the first principal component. The intuition is that a subtask that is not well correlated with the other subtasks may be measuring something different from the intended underlying construct – this subtask should therefore carry less weight in a composite index.

The preceding analysis has suggested that learning performance is balanced across treatment regimes. Table 15 shows results based on regression analysis to test for balance – to test if the differences in average scores in learning outcomes between treatment groups are statistically significantly different from zero. Each column shows a separate regression on treatment indicators after controlling for district and strata fixed effects. The standard errors are clustered at the school level. One star indicates that the difference in means between one of the treatments and the control is statistically significant at the 10% level. The bottom three rows show the p value for the equality tests on the treatment coefficients. In other words, it shows the pair-wise tests comparing the means between treatment groups. A p value less than .05 would indicate imbalance between the respective treatment groups for the relevant learning outcome. The samples are clearly balanced. Out of the 42 possible comparisons, there is slight imbalance in only 2 cases.

Table 15: Balance tests

	(1) Vocab	(2) Letters correct	(3) Digit span	(4) Phon. awareness	(5) Words correct	(6) Compr. test	(7) Sentence	(8) Combined
Training	-0.0367	1.220	0.440	-0.162	-0.0349	0.121	-0.0164	0.0468
	(0.13)	(1.03)	(0.30)	(0.39)	(0.78)	(0.20)	(0.58)	(0.15)
Coaching	-0.203	1.211	-0.120	-0.100	-0.250	-0.156	-0.801*	-0.109
	(0.15)	(1.51)	(0.30)	(0.40)	(0.72)	(0.18)	(0.45)	(0.15)
Parents	-0.119	0.338	0.230	-0.0772	-0.655	-0.128	-0.681	-0.105
	(0.15)	(0.87)	(0.29)	(0.39)	(0.54)	(0.18)	(0.43)	(0.13)
Control Mean	8.658	4.486	4.873	2.246	2.114	0.766	1.546	0.038
Obs	4539	4452	4539	4539	4447	4539	4539	4385
Training=Coaching: p-value	0.258	0.996	0.079	0.878	0.799	0.191	0.157	0.328
Training=Parents: p-value	0.565	0.383	0.490	0.832	0.377	0.236	0.223	0.295
Coaching=Parents: p-value	0.609	0.558	0.256	0.955	0.516	0.887	0.766	0.979

Note: Each column represents a separate regression on treatment dummies and stratification dummies. Standard errors are clustered at the school level. Bottom three lines show p value of equality of coefficient tests between each treatment dummy

PUPIL CHARACTERISTICS

Table 16 plots results for basic pupil-level characteristics. The average age for pupils is 6.37 years and 47% of the sample is female. The median age is also 6, although a sizable proportion (13%) is 5 years old and roughly 9 percent are older than 7.

Table 16: Descriptive statistics – learner age and gender

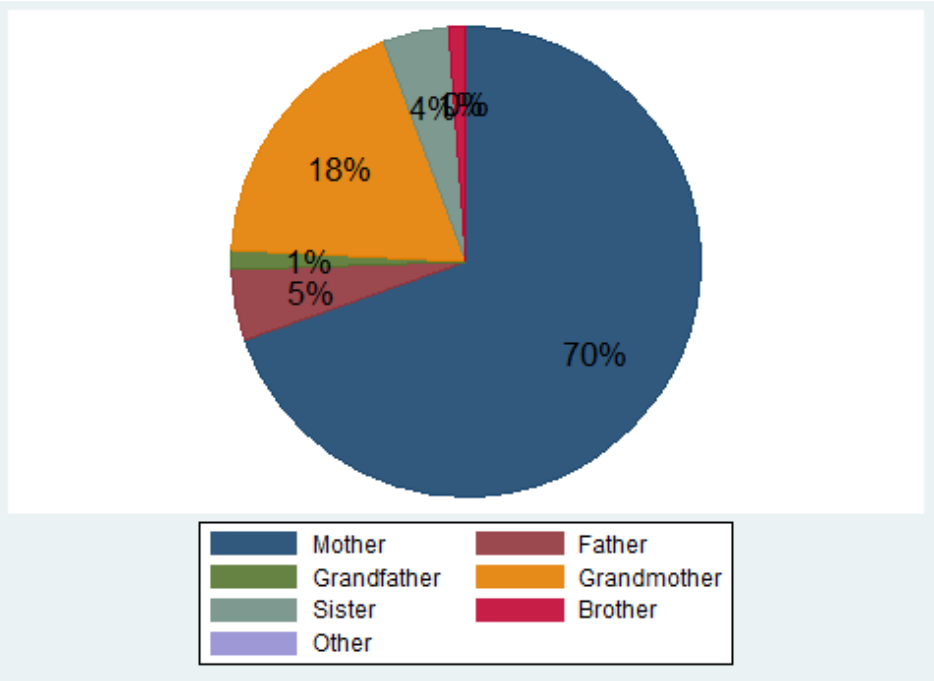
	Observations	Mean	Std. Dev.	Min.	Max.
Pupil age	3881	6.47	0.70	4	10
Girl	4198	0.47	0.50	0	1

PARENT CHARACTERISTICS

Next, we turn to parent characteristics. We sent a survey home with the pupils for the parents/guardians to fill in and return to the school. We only received the forms from 2,484 parents (out of a total of 4,539 pupils who were tested), from only 181 schools. The high response rates in some schools suggest that it is possible to require parents to complete the form. The fact that data is completely missing in 49 schools means that this was a problem of data collection and enumerator training, which should be improved on in the midline and endline rounds of data collection.

Figures 17 to 23 show the main results in a bar graphs and pie charts. First, we discuss parent characteristics. In 97% of cases the primary caregiver filled in the form (not shown). The median age is very young, roughly 25 (note that many gave an answer of 6 or 7 for each. They clearly answered the child's age and we excluded that from the age sample). It is mostly the mother who fills in the form, but note that in a sizable portion (19%) of cases it is the grandmother or grandfather that fills in the form. In only 5% of cases did the child's father complete the form. Even when we restrict the sample to those who claim to be the primary care-givers, roughly 20% are grandparents and over 5% are siblings. However, the mean age of the siblings is 22 years (median is 26), so these are mostly adults.

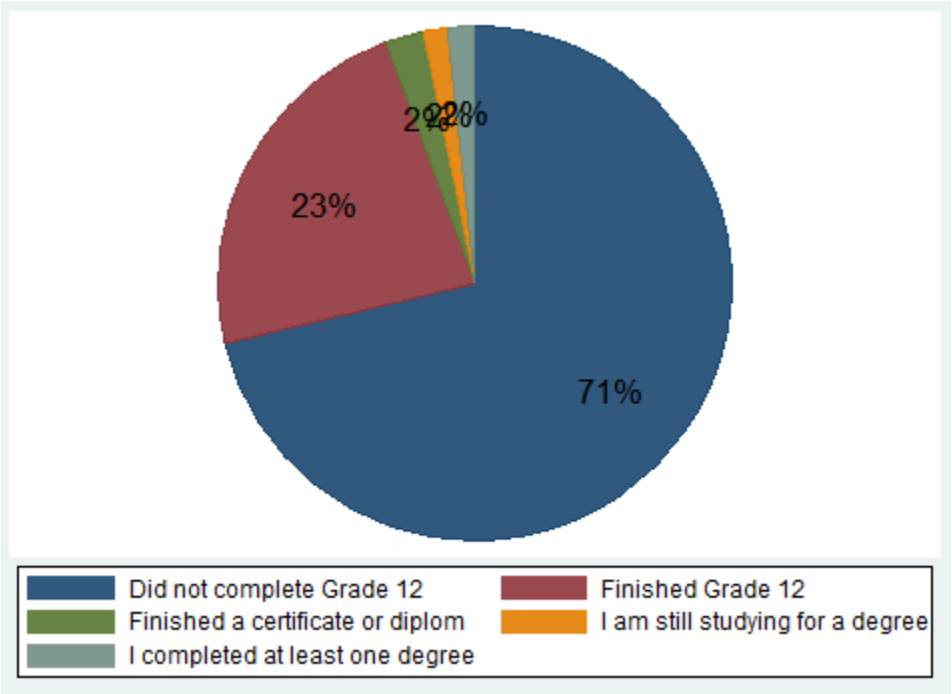
Figure 17: Relationship to pupil



We can see from Figure 18 that most parents have only low levels of education – 71% did not finish matric. Less than 10% completed have a post-matric degree. This size is slightly larger for

grandparents (85% did not finish matric). Unsurprising, given the levels of education, the caregivers rarely read for their own pleasure (not shown). Over a third read less than an hour per week; 41% read 1 to 2 hours.

Figure 18: Education of Guardian



Next we discuss reading activities at home and parents' beliefs and aspirations. Only 10% acknowledge that they never read to their child, yet 27% don't have any books at home and over a quarter read less than one hour a week for their own pleasure. The majority of parents claim that they check if the child is doing his/her homework daily. A third of parents had not spoken to teachers (but since fieldwork occurred in February we cannot place too much weight on this). Only 57% know when the most recent School Governing Body (SGB) meeting was held. This all suggests that there was no parent-teacher meeting at the beginning of the year in many schools or that it was not well communicated to parents.

Figure 19: How often do you read to your child?

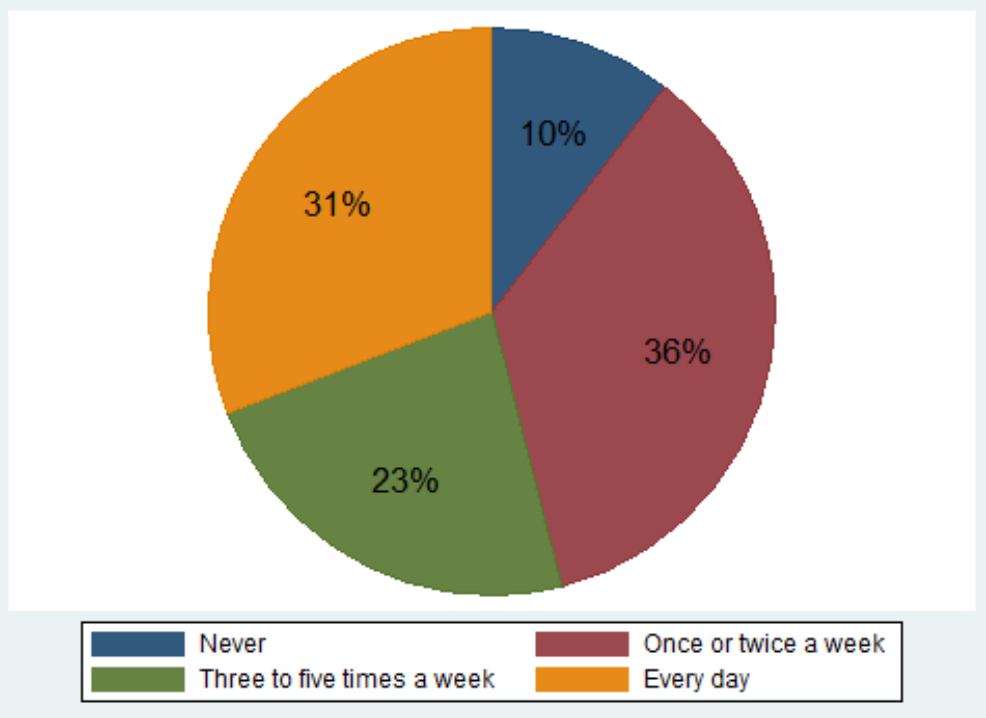


Figure 20: Number of books at home

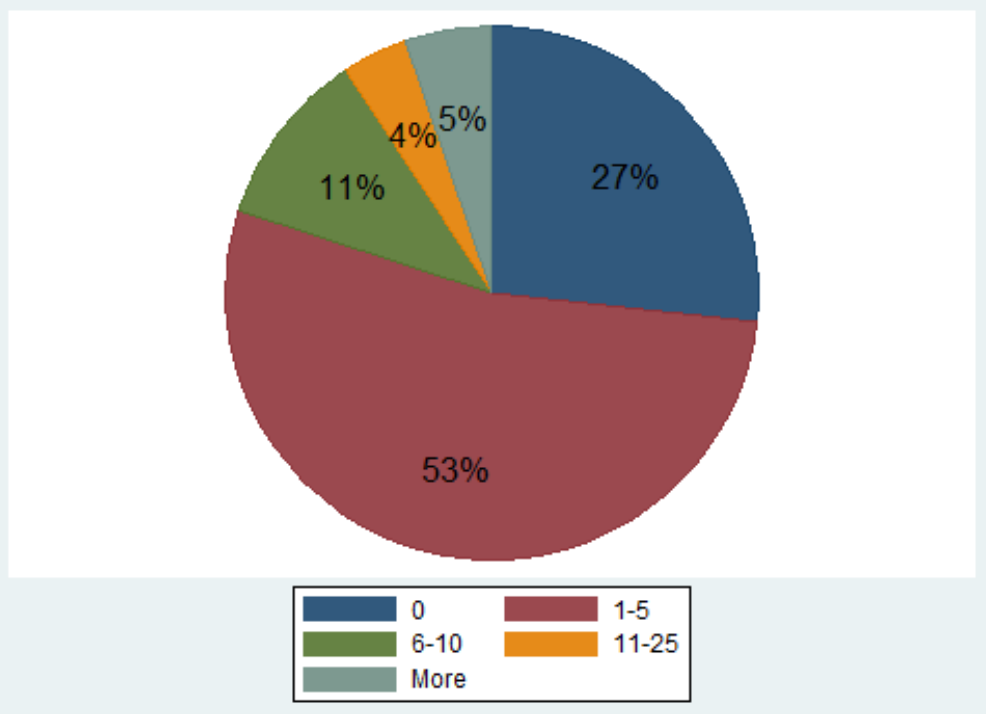


Figure 21: Number of hours read for own pleasure

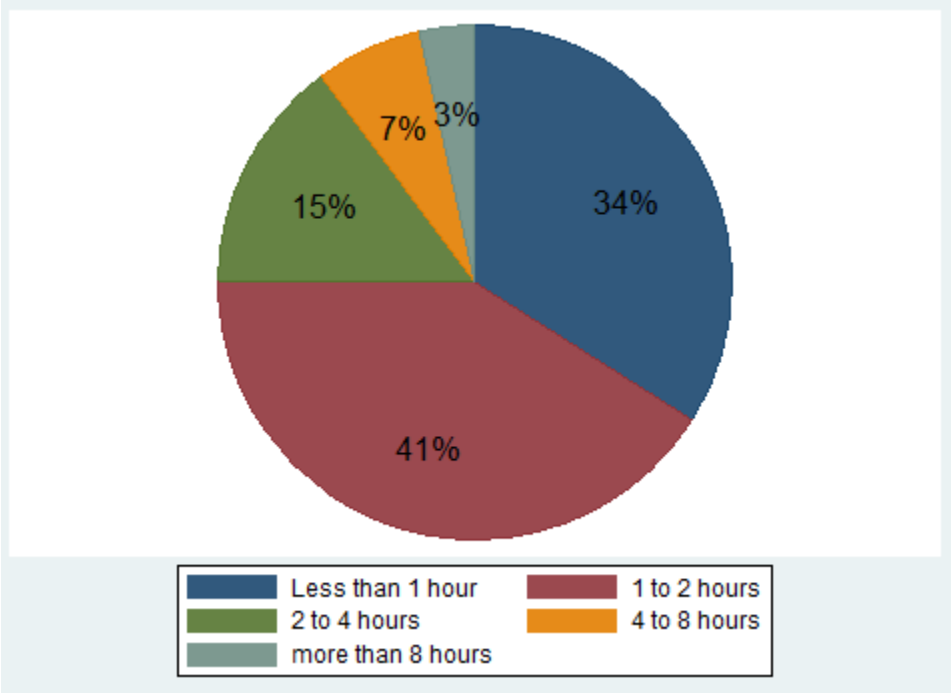
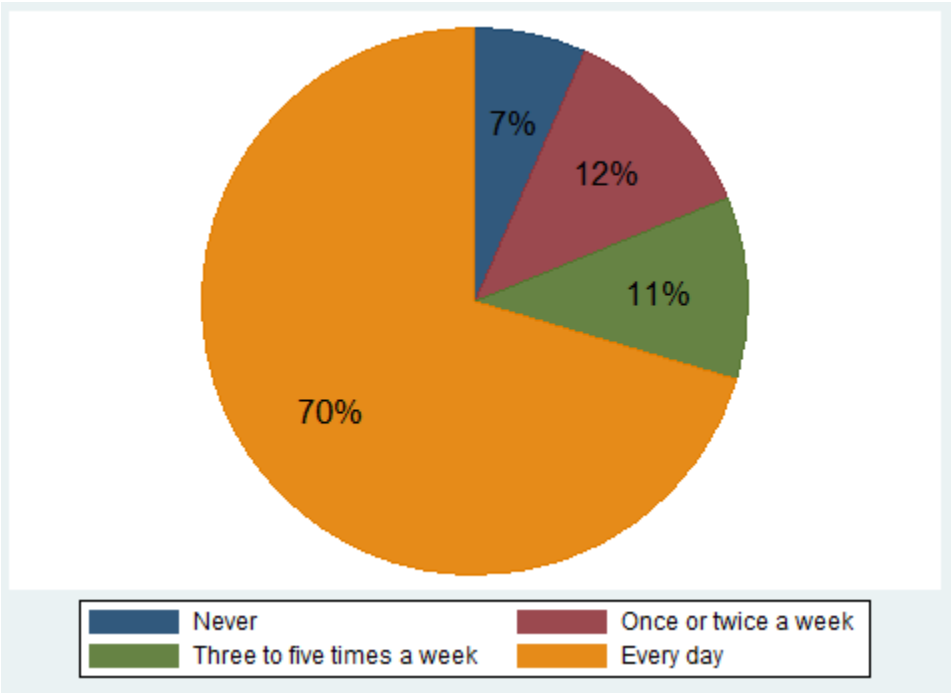


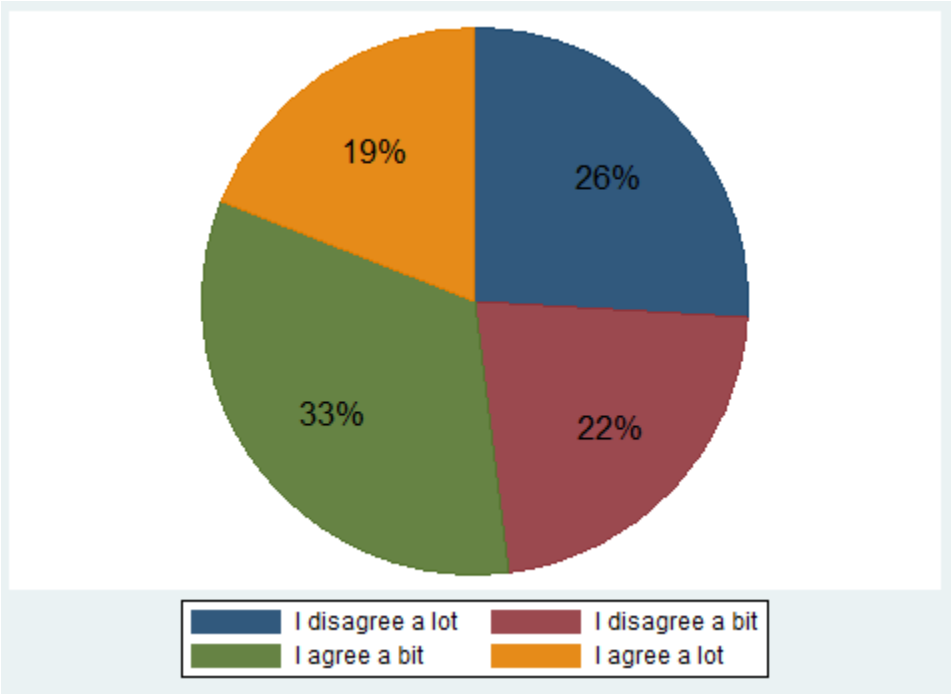
Figure 22: Check that child is doing homework



Just more than half of parents believe that learners at their child's school read poorly. This shows that many parents are critical of the quality of education on offer but a substantial

proportion are probably underestimating the extent of the problem of low learning in schools. It will be interesting to monitor parent beliefs about this at midline and endline assessments, especially in Treatment 3 schools, which are receiving the parent involvement intervention.

Figure 23: Agreement with the following statement: "Our school's learners read very poorly."

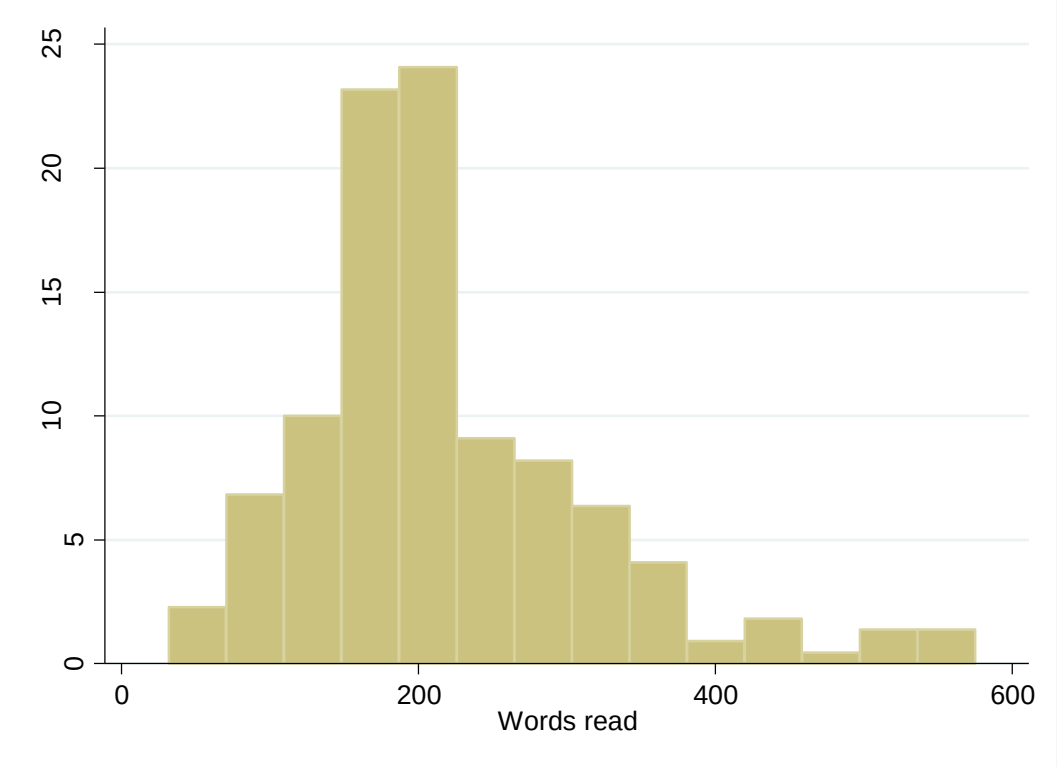


TEACHERS

TEST PERFORMANCE

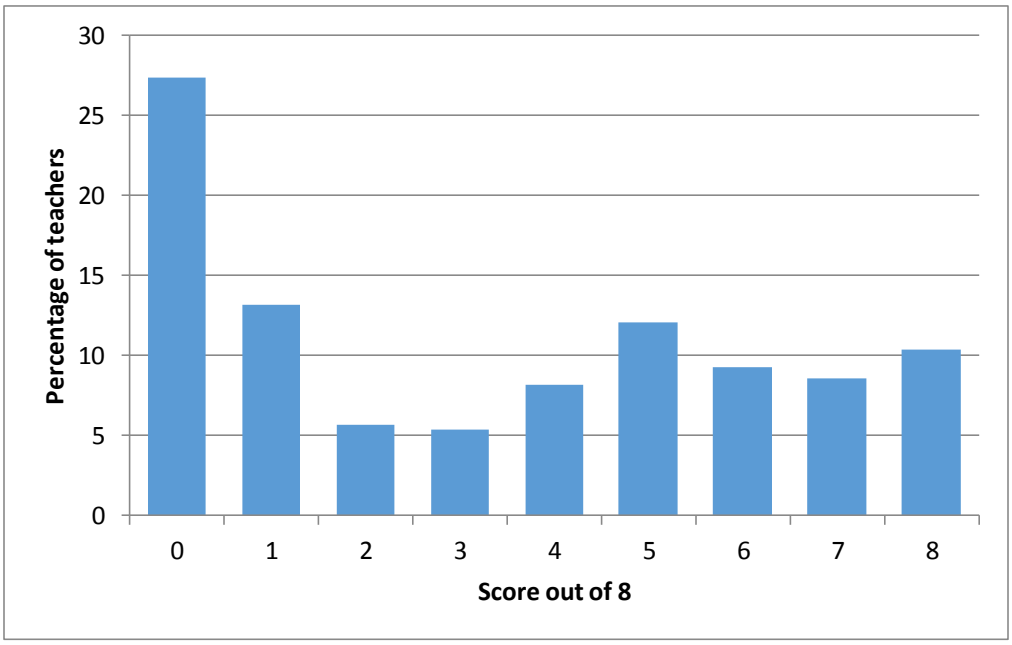
We asked teachers to fill in a questionnaire by themselves whilst the fieldworkers tested the pupils. The teachers were also asked to participate in a short Setswana reading fluency and comprehension assessment. As discussed earlier, there is quite a bit of missing data on the teacher assessment due to non-return of forms in some schools and refusal to participate by 40 teachers. We are left with teacher test data corresponding to about 70% of learners. In the first component of the teacher assessment teachers were given 60 seconds to silently read through a Setswana text consisting of 575 words. The teacher was then asked to indicate how far he/she had gotten. Figure 24 indicates that there is a relatively normal distribution in the number of words read in 60 seconds. The majority of teachers claimed to read a between 150 and 250 words. These estimates of fluency are probably slightly generous but at least they are not wildly unrealistic.

Figure 24: Words read by teachers in 60 seconds



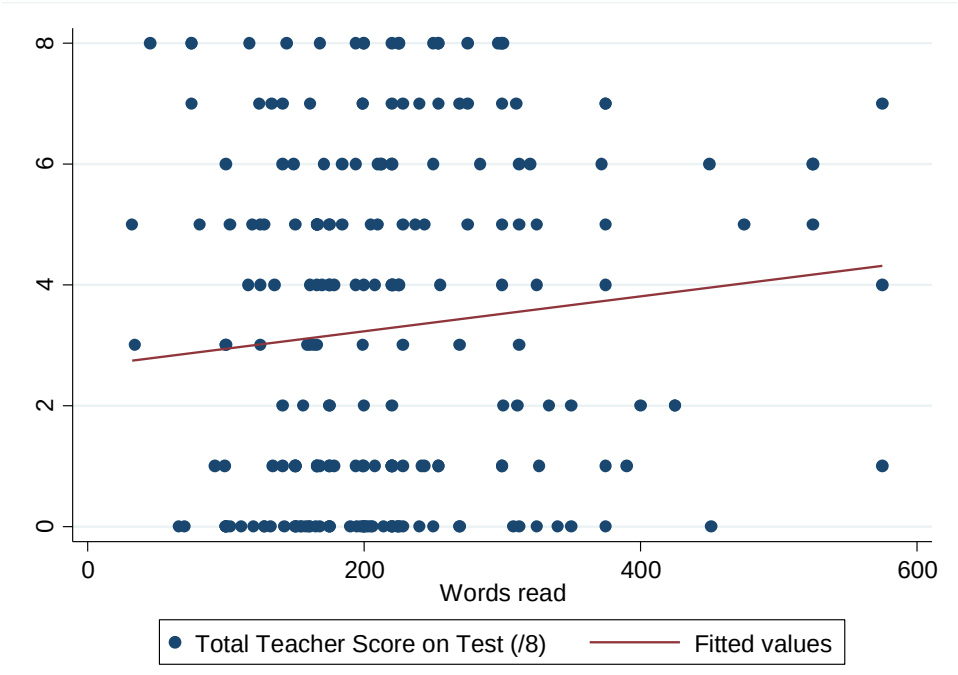
The second section of the teacher assessment was a comprehension test. After spending 60 seconds reading through the text, the teacher was then handed a set of eight multiple choice comprehension questions based on the text. The teacher was then given another three minutes to complete the comprehension questions. The time limit was imposed so as to test the fluency with which teachers are able to read through the text and retrieve answers. The results are disappointing. About 27% of teachers did not get a single question right (Figure 25). It is also possible that these results are biased upwards, if the more competent teachers were more likely to agree to take the test.

Figure 25: Scores on the teacher comprehension test



Furthermore, we can see from Figure 26 that there is a positive correlation between the number of words read and performance in the comprehension test, but this relationship is not nearly as strong as in the case of the pupil test. The weaker validity of the results suggested that administration of the teacher tests may not have been very consistent.

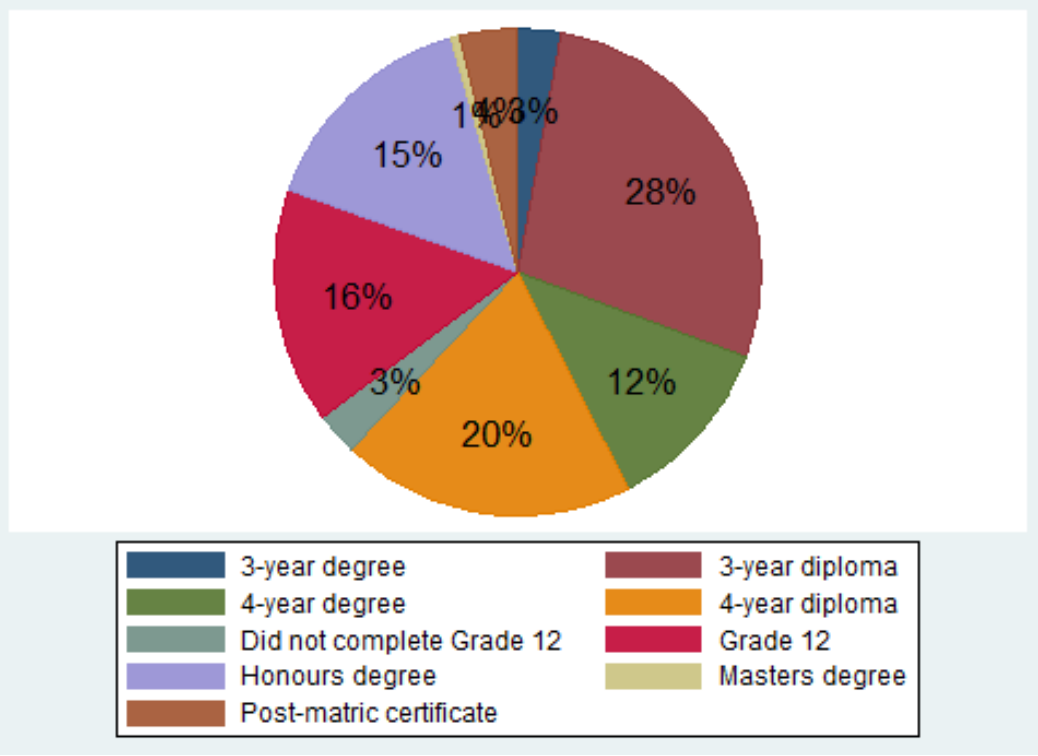
Figure 26: Relationship between words read and questions answered correctly



TEACHER CHARACTERISTICS

Next, we present basic teacher characteristics, shown in Figures 27 to 29 and Table 17. Figure 27 shows the distribution of teachers' education level (note that 45% of teachers did not answer this question. So it is hard to have any confidence in this question). Most have at least a 3-year diploma. 15% only have matric qualifications; 3% have not completed matric.

Figure 27: Teacher qualifications



Figures 28 and 29 reveal an interesting discrepancy in teachers' beliefs. The majority of teachers believe that children should be able to read Setswana fluently (a passage of 50 words in a minute with comprehension) by the end of grade 3 or earlier (84%), yet a slim group actually believe that all children in their school could read by the end of grade 3. Almost half of teachers estimated that only 50-75% of children in their school are able to read by the end of grade 3; roughly a third expect less than half will be able to read! This result is exactly the same if you restrict the sample to teacher who stated they expect that pupils *should* be able to read by grade 3. So, in general teachers don't believe that pupils will reach their expectations.

Figure 28: At what grade should pupils be able to read fluently?

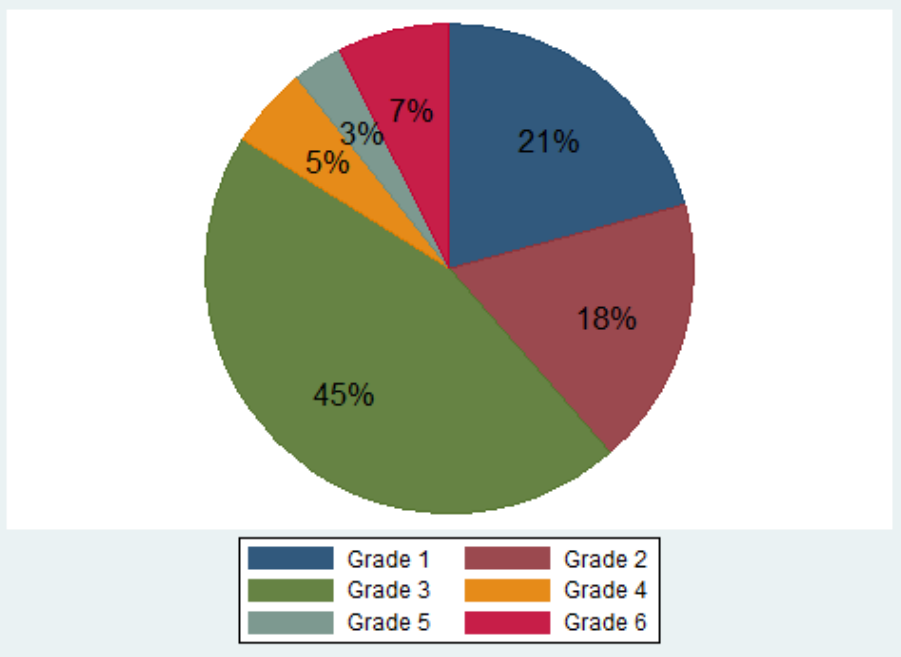


Figure 29: What proportion of children will be able to read by grade 3?

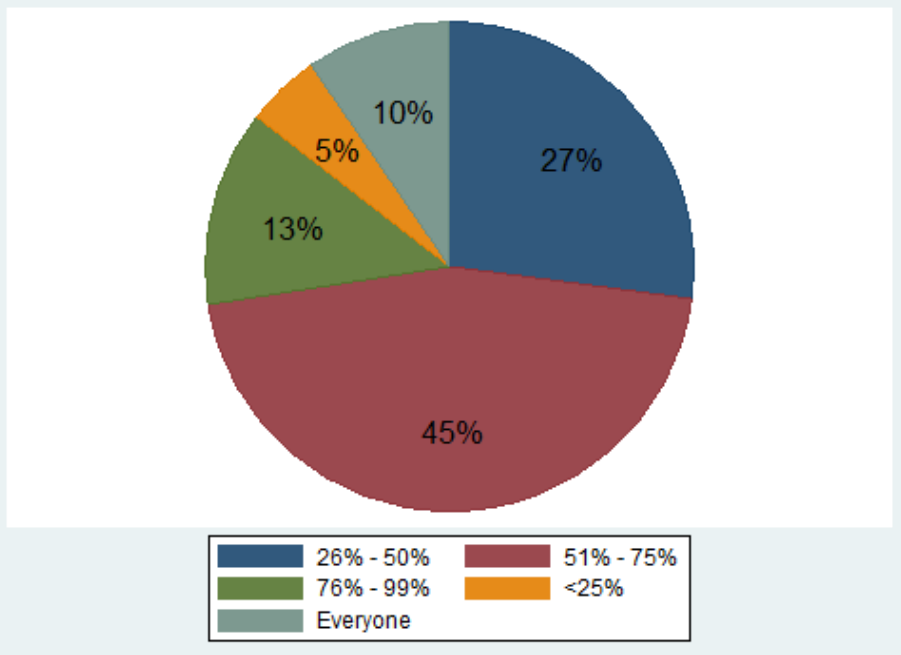


Table 17 shows basic teacher characteristics. Teachers are almost always female, and are rather old with an average age of 50. About 26% of teachers are 56 years old or older. Less than 10% of teachers are younger than 40. 92% speak Setswana most often at home. The average days missed (absent) over the last 10 school days is 1.3; only 49% of teachers claimed not to have been absent at all over the preceding 10 school days. The average class size

("pupils enrolled") is 41 learners. In the majority of classrooms, all the pupils have workbooks and CAPS books. In 60% of classrooms, all the pupils have graded readers (but note the number of missing values for this question). It is unfortunate that on many items there is a lot of missing data due to both non-return of instruments and item non-response.

Table 17: Selected descriptive statistics – teacher characteristics

	Observations	Mean	Std. Dev.	Min.	Max.
Time spent per week teaching reading in Setswana (minutes)	262	171.53	114.50	2	480
days missed last 10 days	203	1.30	2.25	0	10
Age	310	49.78	7.87	24	63
Class size	315	40.62	10.03	3	80
Female teacher	324	0.99	0.10	0	1
Speak Setswana most often at home	311	0.92	0.27	0	1
All learners have CAPS workbooks	241	0.71	0.46	0	1
All learners have workbooks	201	0.78	0.41	0	1
All learners have graded readers	178	0.60	0.49	0	1
At least 3-year diploma	179	0.78	0.42	0	1
At least 3-year degree	179	0.30	0.46	0	1

SCHOOL AND SCHOOL PRINCIPAL

Next, we discuss results from the school principal survey, as shown in Table 18 and Figures 30 to 33.

The majority of school principals have an honours degree, with an average age of 51. The schools are mostly remote rural, with only 1% in a formal suburban area. The schools also come from areas with low levels of socio-economic status: over half of school principals estimated that less than 20% of households have both parents employed. More than half of school principals estimate that the majority of parents have not completed secondary school. The pupil-teacher-ratio in the foundation phase is 38. Almost all the schools (93%) have Setswana as the first language and have a formal language policy.

Table 18: Selected descriptive statistics – school principal questionnaire

	Observations	Mean	Std. Dev.	Min.	Max.
2(a) Age of principal at last birthday:	206	50.85	6.45	28	65
2(c)(vi): Number of filled FP teacher posts	205	5.64	3.14	1	18
Number of foundation phase pupils	230	186.85	134.02	0	545
PTR in Foundation Phase	205	37.84	18.30	0	136
Setswana offered as LoLT	230	0.93	0.26	0	1
School has Formal Language Policy	208	0.91	0.29	0	1
4(a): Learners receive daily meal	230	0.90	0.29	0	1

Figure 30: School principal highest level of education

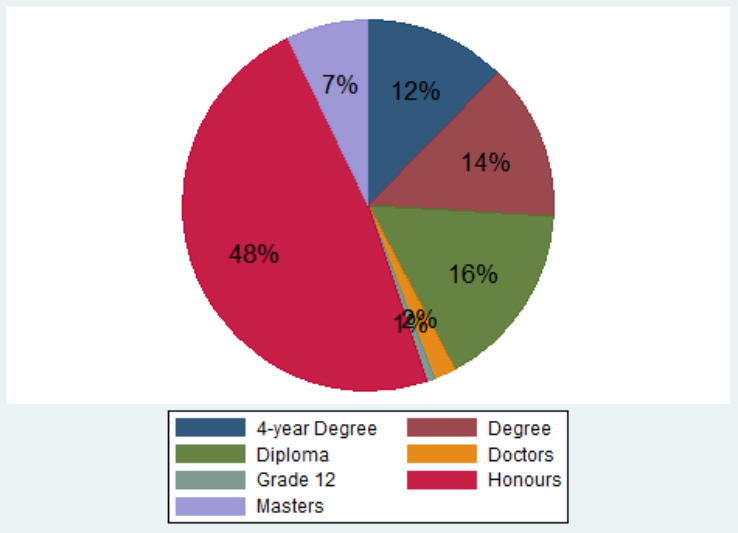


Figure 31: School principal's estimate of average level of parent education

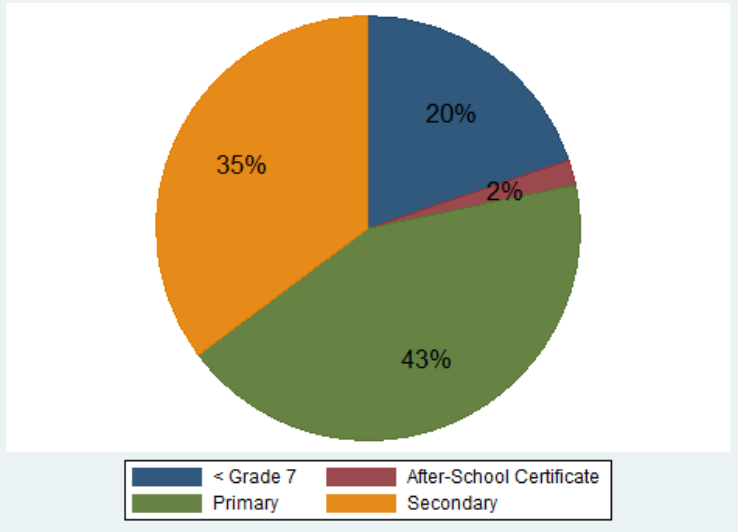


Figure 32: Proportion of learners with both parents employed according to school principal

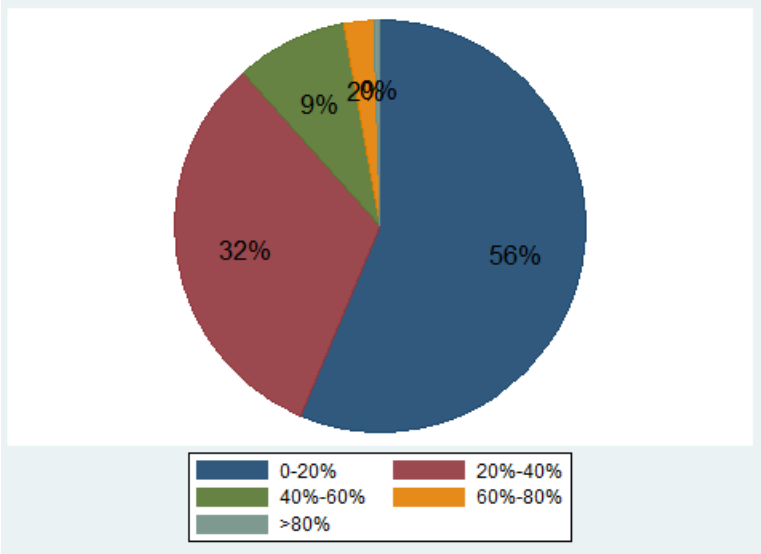
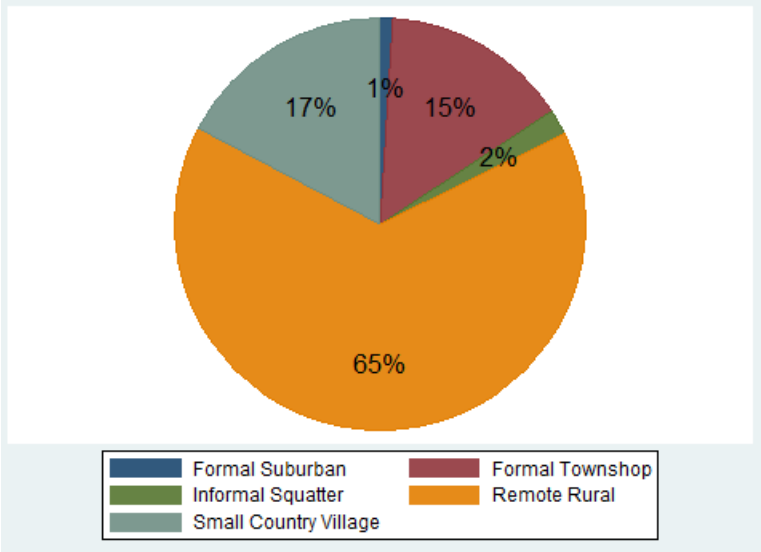


Figure 33: Location of the school according to school principal



WHAT PREDICTS PUPIL TEST SCORES?

Next we examine some pupil, teacher and school characteristics that predict pupil learning outcomes. Note that we can make no causal claims with these regressions – this is merely descriptive analysis, which is informative of possible trends and again tests the validity of our test instruments.

PUPIL CHARACTERISTICS

Table 19 shows results from simple OLS regressions predicting overall composite test score (in terms of standard deviations) (columns 1-3) and letters correct (columns 4-6). The regressions indicate that age does not significantly predict performance. Girls, however, performed significantly better than boys. Girls were able to read about 1 letter more than boys on average, and performed just less than 0.09 SD higher than boys on the composite score. This advantage for girls is consistent and of a similar magnitude with what is observed in standardized tests for higher grades in South Africa, such as in the Annual National Assessments (DBE, 2014), in grade 4 (Howie *et al,* 2012; Fleisch *et al*, 2015) and in grade 6 (Spaull and Taylor, 2015). It is interesting that this gap is evident right at the start of grade 1, which would suggest that the disadvantage may be due to some factor other than school practices that favour girls, most likely differences in the physiological development of girls and boys at this age. Table 19 also indicates that learner performance was significantly better in the district of Ngaka Modiri Molema than in the district of Dr Kenneth Kaunda. This was not anticipated since neither district has shown consistently higher performance in the Annual National Assessments since 2012. The difference between districts will continue to be monitored in the midline and endline assessments.

Table 19: Performance by district, learner age and gender (OLS Regressions)

	(1)	(2)	(3)	(4)	(5)	(6)
	Ave. score	Ave. score	Ave. score	Letters correct	Letters correct	Letters correct
Pupil age	-0.0327		-0.00387	1.078***		1.135***
	(-0.84)		(-0.10)	(3.37)		(3.36)
Female		0.0877***	0.110***		0.992***	1.021***
		(2.77)	(3.44)		(3.11)	(3.71)
District Dummy			-0.456***			-0.784
			(-5.19)			(-1.08)
Observations	3742	4055	3447	3800	4115	3498
Adjusted R^2	0.000	0.002	0.042	0.007	0.002	0.011

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

SCHOOL AND TEACHER CHARACTERISTICS

Table 20 reports regression of school characteristics on the composite reading proficiency score (columns 1-3) and letters correct (columns 4-6). Note first that the school principal's level of education does not matter. This is not too surprising, since these children just joined grade one. Yet, we see that the official poverty quintile of the school (which reflects community level poverty) matters: pupils from schools that are classified as falling in the lowest quintile in terms of socio-economic status perform worse than those in quintile 2. The difference in outcomes

between quintile 1 and quintile 3 is also large in magnitude, although not statistically significant at conventional levels. These trend holds, after controlling for district and location. Unsurprisingly, rural schools also perform worse.

Table 20: Performance by school principal, location, and socio-economic background

	(1) Ave. score	(2) Ave. score	(3) Ave. score	(4) Letters correct	(5) Letters correct	(6) Letters correct
Principal - higher degree	0.0437		-0.0467	-0.508		-1.079
	(0.39)		(-0.38)	(-0.70)		(-1.36)
Quintile 2		0.475***	0.293*		3.442**	1.746**
		(3.06)	(1.97)		(2.52)	(2.08)
Quintile 3		0.130	0.0780		1.126	1.511
		(1.19)	(0.57)		(1.56)	(1.50)
Rural			-0.298*			-1.189
			(-1.95)			(-1.54)
District Dummy			-0.470***			-0.441
			(-4.84)			(-0.56)
Constant	-0.0681	-0.166***	0.207	4.906***	3.840***	5.306***
	(-0.81)	(-3.24)	(1.16)	(8.23)	(11.43)	(5.79)
Observations	2975	4385	2762	3018	4452	2805
Adjusted R^2	0.000	0.040	0.086	0.001	0.021	0.015

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 21 reports regression outputs of teacher characteristics on pupil performance. Surprisingly, teacher performance in the knowledge test is positively correlated with pupil performance, even after controlling for community characteristics: school quintile, location, district, as well as the randomisation strata. It is unlikely that this reflects a causal relationship between teacher quality and learner performance since learners have just joined the school. It is possible that this reflects a selection effect where both learners and teachers select themselves into better schools. Teacher age and education do not predict performance.

Table 21: Performance by teacher characteristics

	(1) Ave. score	(2) Ave. score	(3) Ave. score	(4) Ave. score	(5) Ave. score
Average Teacher Score	0.301***	0.238**	0.355***	0.355***	0.309**
	(3.45)	(2.07)	(3.82)	(3.82)	(2.37)
Age		0.00514			0.00407
		(1.11)			(0.75)
At least 3-year diploma		-0.0871			-0.0814
		(-0.88)			(-0.75)
Rural			-0.0104	-0.0104	-0.106
			(-0.12)	(-0.12)	(-0.76)
Quintile 2			0.0653	0.0653	0.0427
			(0.74)	(0.74)	(0.35)
Quintile 3			0.0778	0.0778	-0.0139
			(0.76)	(0.76)	(-0.10)
District Dummy					-0.182
					(-1.27)
Constant	-0.287***	-0.459*	-0.334***	-0.334***	-0.308
	(-7.56)	(-1.83)	(-3.72)	(-3.72)	(-0.90)
Observations	3077	1395	2787	2787	1259
Adjusted R^2	0.023	0.020	0.034	0.034	0.045

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

PARENT CHARACTERISTICS

Table 22 reports regression results of guardian characteristics on pupil test scores. As before, each column represents a different regression and standard errors are clustered at the school level. Figures 34 to 37 show the main results graphically.

These results show clearly that the home environment matters greatly. Figure 34 shows that pupils do worse if their guardian has not completed matric (there is strangely also a negative result for degree, but a very small sample gave this response). In figure 35 we can see that pupils did far worse in homes where the guardian reportedly never reads to their child; similarly for homes where the guardian reports to never check homework. Figure 36 shows that pupils do better in homes with many books. Even more interestingly, from Figure 37 we see that for

parents that believe they are responsible (and not government or teachers) their child performs far better.

There is some evidence that a child does worse if his guardian is his sister (not shown), possible because these are orphaned households. But the small sample means we shouldn't place too much value on this result. There is also suggestive evidence that for pupils whose guardians believe their child can improve in learning performed better, but this result is not strong.

Table 22: Parent characteristics and learner performance

	(1) Ave. score	(2) Ave. score	(3) Ave. score	(4) Ave. score	(5) Ave. score
Finished matric	0.248***				0.224***
	(3.66)				(3.27)
Degree	0.277				0.314
	(1.39)				(1.31)
Reads to Child		0.229***			
		(2.82)			
Homework every day		0.0836*			0.0574
		(1.67)			(1.10)
Believe - responsible for learning			0.265***		0.0805
			(3.46)		(1.32)
Books				0.0124***	0.00830*
				(2.91)	(1.85)
District Dummy					-0.317***
					(-3.40)
Quintile 1					-0.327***
					(-2.78)
Constant	-0.0554	-0.244***	-0.244***	-0.0550	0.00833
	(-0.94)	(-3.18)	(-4.90)	(-0.89)	(0.08)
Observations	2080	2043	2212	2090	1728
Adjusted R^2	0.011	0.006	0.004	0.003	0.061

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < .01$

Figure 34: Parent education and learner test scores

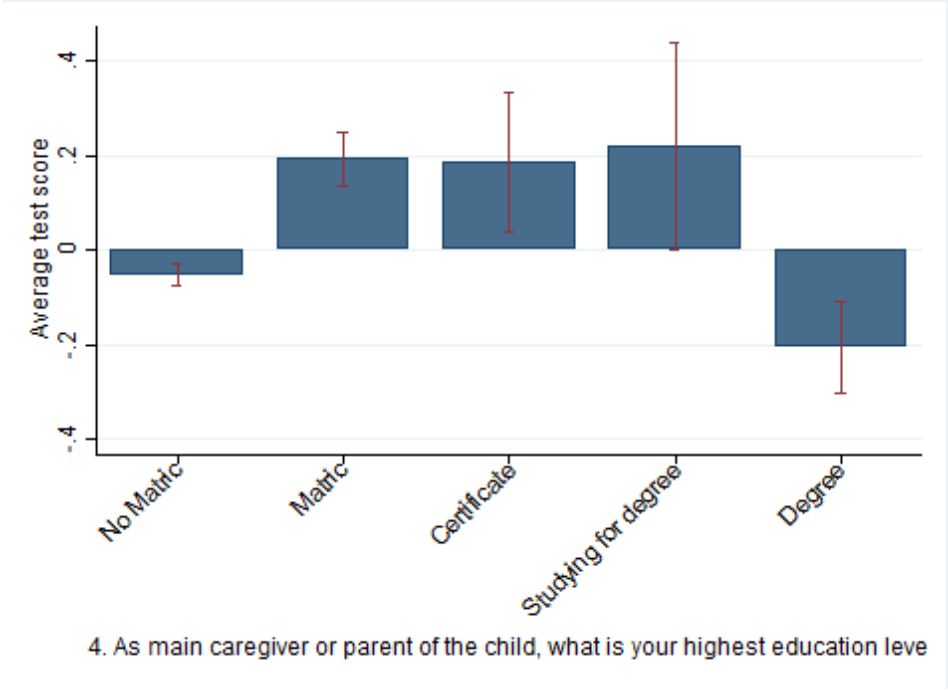


Figure 35: Parent reading to child and learner test scores

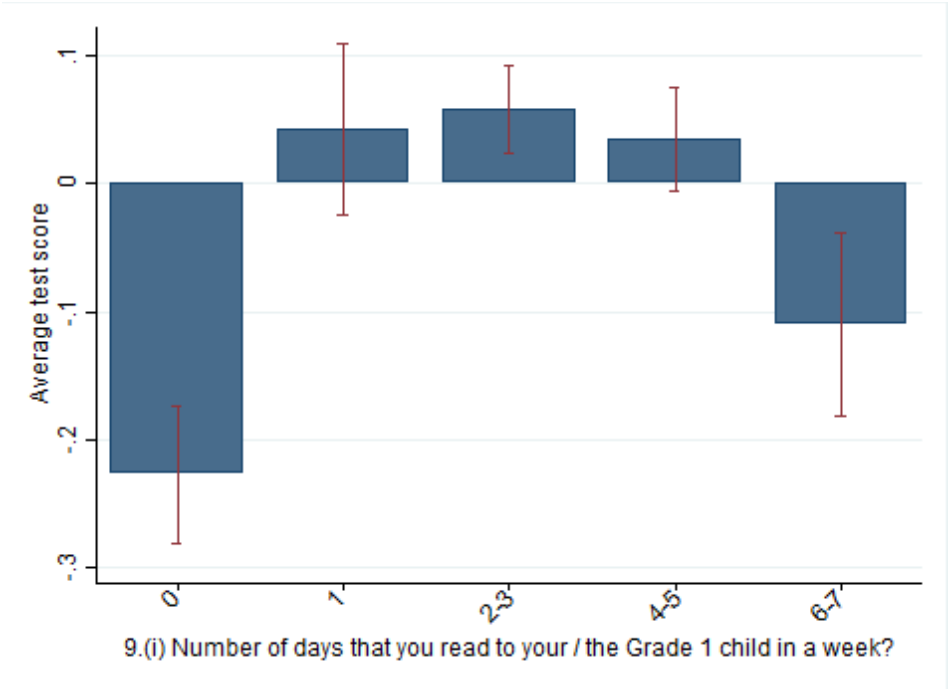


Figure 36: Parent checking homework and learner test scores

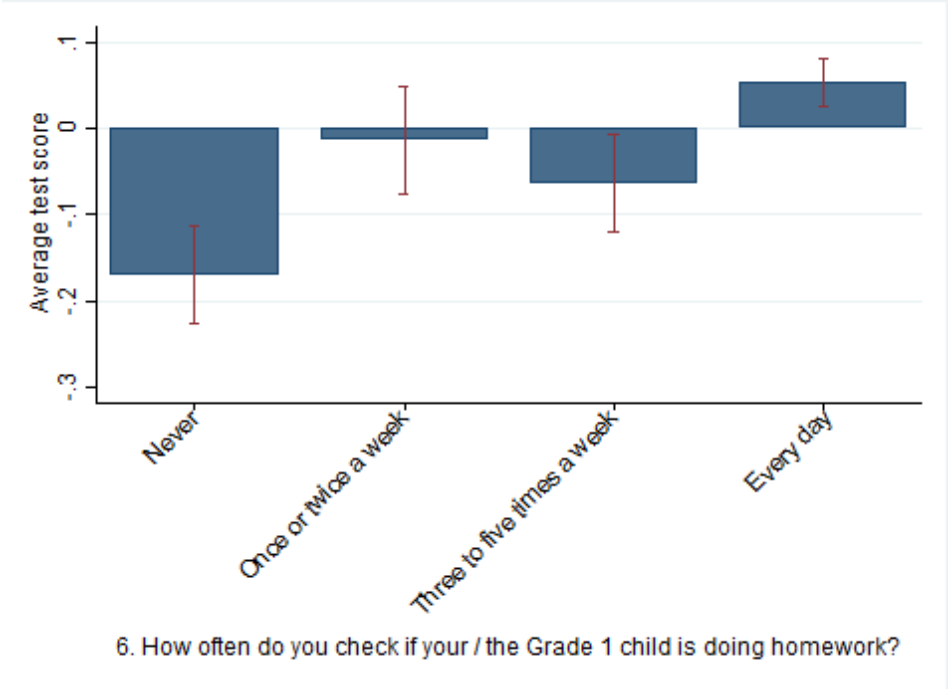
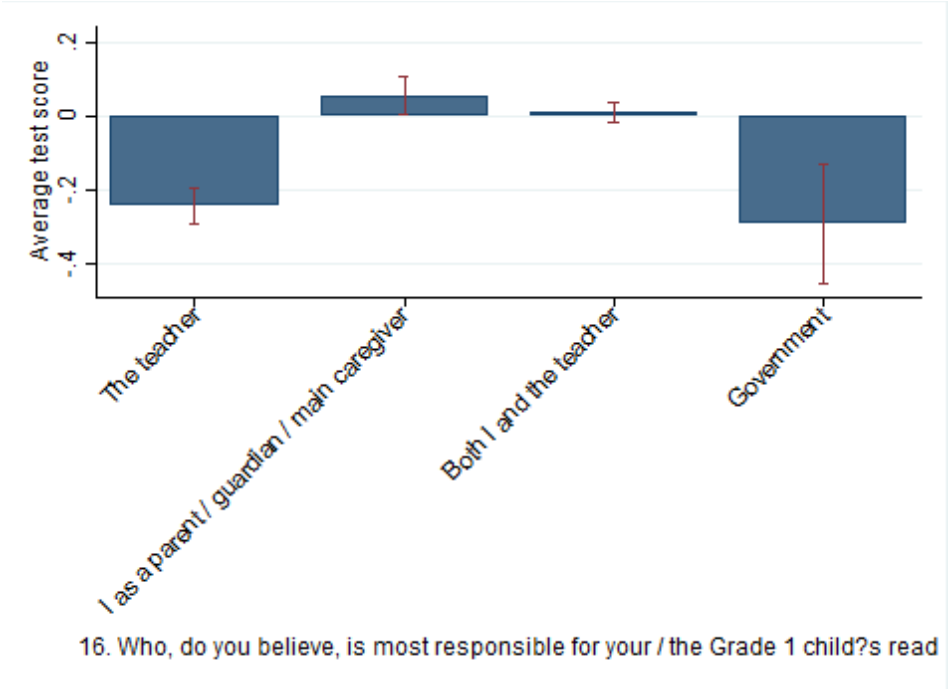


Figure 37: Parent's perceived responsibility for learning and learner test scores



Reading for own pleasure doesn't appear to matter. The following factors also do not seem to be correlated with test scores: reported frequency of meeting with teacher; attitudes regarding the importance of Setswana, beliefs about their child's learning ability, beliefs about the quality of the school.

To summarize: parents' education and involvement in their child matter. Involvement in the school or their own beliefs over learning and their child's ability doesn't matter. A sense of agency – a belief that they are important to their child's learning – does matter. All told, the self-reported home-level characteristics – guardian's education level, beliefs, and involvement in their child's reading - are strongly correlated with pupil test scores. A guardian's involvement in the school (attending SGB meetings etc.) doesn't seem that strongly correlated with reading scores.

PROGRESS REPORT ON IMPLEMENTATION OF INTERVENTIONS

EGRS Treatment 1 (training)

Treatment 1 trains the teachers on how to use the lesson plans and accompanying materials through central training sessions, each lasting 2 days, and occurring twice yearly. The first session was conducted in February 2015 and the second occurred in July 2015. Similar sessions are scheduled for 2016.

SUMMARY OF KEY EVENTS

Reference Group Meetings

Two reference group meetings were held with Provincial Officials from the Quality Assurance and Research Directorates, and with Foundation Phase Language specialists working in the Dr Kenneth Kaunda District and the Ngaka Modiri Molema District education offices. The purpose of these reference group events was primarily to ensure that the lesson plans and support materials are aligned to existing curriculum support offered in the province but also to introduce the group to the EGRS programme and approach, to garner buy-in for the programme; to critically engage with the materials used in order to strengthen them and to reflect on successes and challenges of the programme and implementation. These events are relevant to both Treatments 1 and 2 since both these treatments make use of the same lesson plans and support materials.

Training Events

- Two training events were held to train school managers and Grade 1 teachers, and to distribute materials.

- The first event was held at the Protea Hotel in Klerksdorp. Approximately half the schools attended the first session, on the 24th and 25th February 2015, and half the schools attended the second session, on the 26th and 27th February 2015.
- The second event was held at the Kedar Country Lodge in Rustenburg. Schools from Kgetleng, Maquassi Hills, Matlosana and Ramotsere Moiloa attended the first session, on the 14th and 15th July 2015. Schools from Ditsobotla, Mafeking and Rekopantswe attended the second sessions, on the 16th and 17th July 2015.
- Accurate, up-to-date data on schools, managers and teachers was gathered at the first training event, allowing for much more efficient logistics at the second event.
- At the first training, teachers were given an overview and technical understanding of the programme, as well as an introduction to classroom management, classroom environment, resources management and core methodologies. This training prepared them to implement the programme at a technical level.
- At the second training, the technical features of the programme were revised, and then teachers were given a more in-depth training on core methodologies, particularly those related to writing. Teachers were also given the opportunity to share the work done in their classrooms.
- At both trainings, school managers were introduced to the concept of supporting and monitoring teachers as they implement the programme. They were also given monitoring and support tools to assist in this process.
- Teachers and managers responded very well to the programme and materials, and it was clear that the second event really deepened the understanding of the purpose of the programme and core methodologies.
- Ongoing challenges related to Treatment 1 include:
    - Non-participation by a small number of teachers;
    - Poor time management skills of some teachers, leading to insufficient curriculum coverage;
    - The limited feedback related to implementation in this model limits the service provider's understanding of teacher challenges.

Table 23: Attendance at Treatment 1 training events

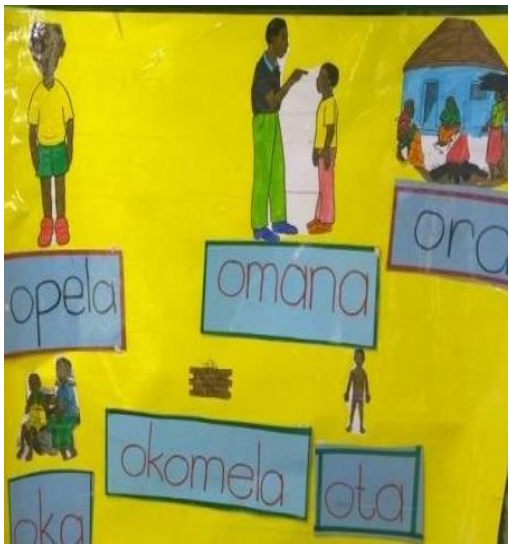
	FEB 2015	JULY 2015
School Attendance at Training	100%	92%
Grade 1 Teacher Attendance at Training	100%	85%
School Leaders Attendance at Training	79%	81%

**Note: Materials are distributed to schools that did not attend training.*

TREATMENT ONE PHOTOGRAPHS



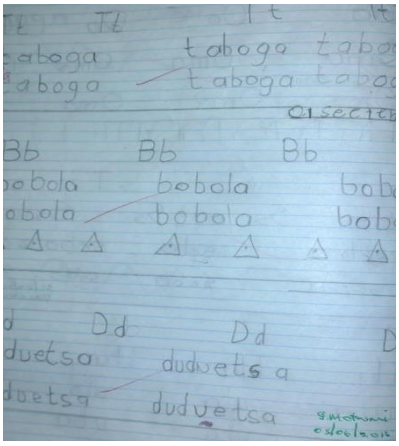
Training: Demonstration by Teacher



Teacher's Chart: Vocabulary



Teacher's Chart: Mind Map



Learner's Work: Handwriting

EGRS Treatment 2 (coaching)

Table 24: Structure of treatment 2

Coach Name	District	Number of Schools	Number of Grade 1 Teachers
Kgomotso Phalatse	Ngaka Modiri Molema	17	26
Helen Kgobane	Ngaka Modiri Molema	18	30
Sabi Mlambo	Dr Kenneth Kaunda	15	34

SUMMARY OF KEY EVENTS

Training

Treatment 2 has 3 full time coaches, each working with a set number of schools and teachers. Coaches have held three teacher training sessions prior to the implementation of the programme for Terms 2, 3 and 4. For Term 2, this was done as 2 x half day sessions, with each coach training 3 – 5 small clusters of teachers. This was not a particularly successful model, as it took a long time to train all teachers, and the smaller groups lacked the energy and enthusiasm of the slightly larger training groups. As a result, the training model was changed after Term 2. For Terms 3 and 4, coaches held full day training sessions to prepare for the next term. Each coach held 2 – 3 larger training events at the end of the previous term, to allow teachers preparation time in the school holidays. In addition, coaches run regular Professional Working Group (PWG) training sessions during the term.

Coaching

During term times, coaches provided follow-up support to all Grade 1 teachers in participating schools. Coaches visit teachers a minimum of once per month. The support sessions include the following kinds of activities:

- Lesson demonstrations by coaches to illustrate the core methodologies
- Lesson observations by coaches
- Critical but positive feedback to teachers regarding lessons observed
- Monitoring of learner exercise and workbooks
- Monitoring of curriculum coverage
- Monitoring of learner assessment results
- Professional interaction with principal and HoDs regarding implementation

Supervision

During term times, supervision of coaches takes place on a regular basis, both on and off site. On site supervision of coaches takes place a minimum of once per term. Treatment Two supervision is characterised by:

- Accompanying coaches on teacher support visits
- Observation of the coaches in practice
- Critical but positive feedback to coaches regarding the manner in which they support teachers
- Informal discussions with teachers concerning the learning programmes and their implementation
- Informal discussions with principals and HODs concerning curriculum and assessment issues
- Monitoring of work schedules and attendance registers

Outcomes of Training and Coaching Sessions

- Successes related to Treatment Two include improvements in:
  - Teacher morale
  - Curriculum coverage
  - Pedagogical content knowledge
- Challenges related to Treatment Two include:
  - Poor-participation by a small number of teachers
  - Slow pacing by some teachers, leading to insufficient curriculum coverage
  - Difficulties related to multigrade teaching in some instances
  - High absenteeism of learners and teachers
  - Practical circumstances – large class sizes and poor infrastructure

Table 24: Attendance at training events

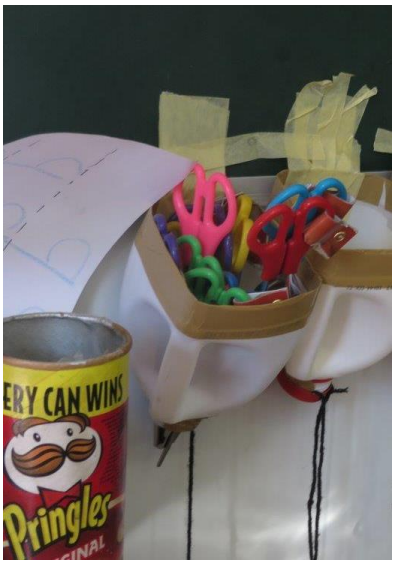
	Term 2	Term 3	Term 4
School Attendance at Training	100%	88%	94%
Grade 1 Teacher Attendance at Training	100%	82%	93%

**Note: 'Catch-up' training sessions are held with teachers who miss the initial training sessions*

TREATMENT TWO PHOTOGRAPHS



Classroom: Word Wall



Classroom: Resource Management



Learner's Work: Writing



Group Guided Reading

EGRS Treatment 3 (parents)

SUMMARY OF KEY EVENTS

Project Launch

Principals and SGB Representatives were invited to a launch event. At the launch event, schools were introduced to the concept of a Community Reading Coach (CRC) and were asked to recruit a CRC for their school.

CRC Training Events

CRCs are regularly trained in four cluster groups: Zeerust; Lichtenburg; Mafeking and Klerksdorp. Training events focus on administration, facilitation skills and pedagogical content knowledge. The topics covered to date are as follows:

- Topic 1: Small Things Make a Big Difference: Getting the Basics Right
- Topic 2: Playing With Sounds to Support Reading
- Topic 3: Reading Pictures
- Topic 4: Letter Sounds
- Topic 5: Incidental Reading
- Topic 6: Preparing to read Books
- Topic 7: Reading Story #1

Family Training Sessions

For each topic, CRCs hold three family training sessions. The same content is covered in slightly different ways over the three sessions. For each topic, families are given a 'family card' with key information to remind them of certain behaviors and practices to implement in their homes. Key information is presented and discussed, and then activities are practiced. Families are encouraged to replicate activities at home on a regular basis.

Outcomes of Family Training Sessions

- Successes related to the family training sessions include:
    - Improvements in the knowledge of parenting and reading support skills of parents and families
    - Improved reading skills of families
    - Slowly increasing social capital of families, leading to better participation in formal school structures
- There are also anecdotal reports of improved school attendance and performance by children of participating parents.

- Challenges related to the family training sessions include:

- Recruitment of suitable CRCs. In some instances, the resident population is either not willing to work for the small volunteer stipend, or there is no suitably skilled candidate available for the position.
- Attendance of families at training sessions is an ongoing challenge. Attendance decreased over the winter months and before and after school holidays.
- New strategies to increase parent attendance include engagement with principals to try and motivate parent attendance, and a small incentive scheme in the form of a 'lucky-draw prize' per session. These strategies were implemented from August, and the results are still to be measured.

Table 25: Attendance at CRC training events

	Orientation & Topic 1	Topics 2 & 3	Topics 4 & 5	Topic 6	Topic 7
CRC Attendance at Training	98%	88%	70%	86%	88%

**Note: 'Catch-up' training sessions are held with CRCs who miss the initial training sessions*

Table 26: Attendance at parent training events

	Orientation & Topic 1	Topic 2	Topic 3	Topic 4
Parent Attendance at Training	35%	29%	21%	18%

TREATMENT THREE PHOTOGRAPHS



CRC Training: Role Play



CRC Training: Games



Family Card 1: Phonemic Awareness



Family Card 5: Incidental Reading



Family Card 7: Reading Books



Family Card 1: The Basics

NEXT STEPS IN THE PROJECT

Interventions are scheduled to continue throughout 2016. A midline data collection is taking place between the 26th of October and the 13th of November 2015. Using this data, the impacts of each intervention after one year will be measured. The endline data collection is scheduled for October/November 2016. This will allow us to measure the impacts of two years of treatment on reading outcomes at the end of grade 2. In the event of at least one of the interventions showing a significant impact on reading outcomes at the end of grade 2, we plan on using DBE administrative test data and possibly even raising funds for a further round of data collection to measure the longer-term impacts of the interventions.

REFERENCES

Andrabi, T. D. (2013). *Report Cards: The impact of providing school and child test scores on educational markets.* Unpublished mimeo.

Department of Basic Education. (2014). *Report on the Annual National Assessment of 2014.* Pretoria: Department of Basic Education.

Fleisch, B., & Schöer, V. (2014). Large-scale instructional reform in the Global South: insights from the mid-point evaluation of the Gauteng Priamry Language and Mathematics Strategy. *South African Journal of Education, 34*(3), 1-12.

Fleisch, B., Taylor, S., Schöer, V., & Mabogoane, T. (2015). The impact of a remedial reading programme on second language grade 4 students in KZN: Evidence from a randomised impact evaluation. *South African Monitoring and Evaluation Conference.* Johannesburg.

Gove, A., & Wetterberg, A. (2011). *The Early Grade Reading Assessment: Applications and Interventions to Improve Early Grade Literacy.* RTI Press.

He, F., Linden, L. L., & MacLeod, M. (2009). *A better way to teach children to read? Evidence from a randomized control trial.* Retrieved September 12, 2012, from https://webspace.utexas.edu/linden/www/Teach%20Children%20to%20Read.pdf

Howie, S., Van Staden, S., Tshele, M., Dowse, C., & Zimmerman, L. (2012). *South African Children's Reading Literacy Achievment: Summary Report.* Pretoria.

Lieberman, E. P. (2013). *Does Information Lead to More Active Citizenship? Evidence from an Education Intervention in Rural Kenya.* MIT Working Paper.

Pandey, P., Goyal, S., & Sundararaman, V. (2009). Community participation in public schools: Impact of information campaigns in three Indian states. *Education Economics, 17*(3), 355-375.

Sailors, M. H. (2010). The Effects of First- and Second-Language Instruction in Rural South African Schools. *Bilingual Research Journal, 33*, 21-41.

Spaull, N., & Taylor, S. (2015). Access to what? Creating a composite measure of educational quantity and educational quality for 11 African countries. *Comparative Education Review, 59*(1), 133-165.

Taylor, S., & Coetzee, M. (2013). *Estimating the impact of language of instruction in South African primary schools: A fixed effects approach.* Stellenbosch Economics Working Papers 21/13.

APPENDIX F: CHAPTER 4 EGRS Letter of Permission

basic education
Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

Private Bag X895, Pretoria, 0001, Sol Plaatje House, 222 Struben Street, Pretoria, 0002, South Africa
Tel.: (012) 357 3000, Fax: (012) 323 0601, www.education.gov.za

Ref no: ODG-2446-29/01/2016
Enquiries: Ms M Matabane
Tel: 012 357 3658
Email: Matabane.a@dbe.gov.za

Ms N Mohohlwane
Chief Directorate: Strategic Planning, Research & Coordination
Department of Basic Education
PRETORIA
0001

By email: Mohohlwane.N@dbe.gov.za

Dear Ms Mohohlwane

REQUEST FOR APPROVAL TO ACCESS EDUCATION DATASETS FROM THE DEPARTMENT OF BASIC EDUCATION, FOR RESEACH STUDY

On 14 January 2016 the Department of Basic Education (DBE) received your research request to access the Early Grade Reading Study (EGRS) Baseline data sets at the DBE; as part of your Master in Education: The Contribution Of Randomised Control Trials (RCTs) to Evaluating System-Wide Early Grade Reading Interventions- Evidence From Developing Countries And South African Case Studies; at the University of Witwatersrand.

The research request is approved on condition that you, as the applicant of the research data,

1. adhere to the conditions set in the research protocol document titled, "Guidelines for researchers in conducting research in the Department" (available on the DBE intranet and internet); and
2. ensure anonymity of the data and maintain confidentiality of sensitive information contained within the EGRS (Early Grade Reading Study) data sets; particularly regarding the EMIS (Education Management Information System) Numbers and School Names information within the data.

Please be aware, you have only been granted approval to access and utilise the EGRS Baseline data sets for this study. Should you require the data for other purposes you will be requested to resubmit a research request to the DBE.

As guided by the Department's Research Protocol, we understand that you are aware of the ethical and legal responsibilities towards the research data and that you will protect the welfare and maintain the privacy and confidentiality of all data records provided to you; at all times.

Please also note, that we request that you share the findings of the research with the DBE at the conclusion of your research. Kindly supply the RCME unit with two copies for the attention of the Director-General.

Yours sincerely

MR HM MWELI
DIRECTOR-GENERAL
DATE: 29/01/2016