The impact of a remedial reading programme on second language grade 4 students in KZN: Evidence from a randomised experiment

Stephen Taylor, Brahm Fleisch, Volker Schöer and Thabo Mabogoane

September 2015

Paper presented at the Economics Society for South Africa conference of 2015

Abstract

The majority of South African children do not speak English as their first language yet are taught in English from Grade 4 onwards. This represents one of the various educational disadvantages that are contributing to the low levels of learning observed amongst the majority of poor children in South Africa. Finding ways to reduce the learning deficits amongst these children is therefore an important policy priority. This paper reports on a randomised controlled trial of a remedial programme designed to boost the English reading and literacy skills of grade 4 students, for whom English is a First Additional Language. The study randomly assigned 100 initially low-performing public schools in the Pinetown district of KwaZulu-Natal to treatment and control groups. The intervention lasted for 11 weeks, was administered within normal school time and consisted of three components: the provision of scripted lesson plans, additional reading resources and on-site instructional coaching for teachers. The 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 programme impact was larger for learners who initially had a basic minimum of English skills and for those whose teachers participated actively in the programme. The paper describes some of the challenges involved in implementing a randomised controlled trial in the context of the South African school system. The paper also reflects on how this sort of impact evaluation presents a challenge to the conventional research in education policy, but also creates valuable opportunities for economists and educationists to collaborate to take knowledge further.

Funding for this study was provided by the Zenex Foundation

Contents

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 acknowledge and thanked for their ongoing support and assistance. We would like to thank the Wits Education Ethics Committee (non-medical) for its scrutiny and flexibility. To Mary Metcalfe, a big thanks 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 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 portion of South African schoolchildren are one or more years below the acceptable achievement levels, particularly in key subjects like English First Additional Language and Mathematics (Taylor, 2014, NEEDU, 2014, Spaull and Kotze, 2015). Spaull and Kotze (2015) 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, i.e. a small number of individuals within a class that have specific learning barriers or challenges, but rather the learning deficits are systemic, often effecting almost all learners across the majority of disadvantaged schools.

How can education departments address these systemic learning backlogs? There are a growing number of specialized programmes, particularly at the Grade 12 levels that focus on providing additional 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 that was developed as a component of the Gauteng Primary Literacy and Mathematics Strategy in 2012. The eleven week programme that focused on re-teaching 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 basic of English language literacy. Hellman's (2012) interval evaluation suggested 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 designed the intervention, the pre and post-instruments were administered by the teachers themselves, and the study did not have an estimate of a counterfactual.

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 Catch-up Programme. Although evidence generated from small scale studies (such as Pretorius, 2014) has the potential to contribute to the knowledge base, there is a clear need for studies with a sufficient sample size and with plausible method for identifying the causal impact to allow policy makers and 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 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 post-tests in

both treatment and control schools. The intervention took place in April to June 2014 in Pinetown, Kwazulu-Natal.

This paper is structured into four sections. Following this brief introduction, the paper 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 post-testing. 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 explores explanations for the main finding. The final section considers the implications of the study.

2. Research Design and Methods

2.1 Background to the Reading Catch-Up Programme

In 2011 the Gauteng Primary Language and Mathematics Strategy developed and implemented an Intermediate Catch-Up Programme to remediate the learning gaps in underperforming Grades 4 to 6 classrooms. The Catch-Up Programme contains three key elements, i.e. scripted lesson plans, provision of high quality learning materials, and on-site coaching. The scripted lesson plans divided the term into 11 weeks, with each week designated with 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 simple 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 is to be sent home every day, and the test book only to be sent home 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 classrooms (Grades 4 to 7) 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 based on research (PIRLS, SACMEQ and ANA) that suggested 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 one set per learner. 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 programme and the graded class readers, what teachers called the 'thin books'. The last learning and teaching resource was a mark book, what the programmed 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 ten sentences using the posters. The second period was for phonics and spelling; two new sounds and related words as well as specific 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 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 around 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, writing and assessment. The assessment took the same form every week, 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 twelve graded class readers, the four posters, and twelve 'look and say' word sheets.

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, day of the week and the date. The lesson time (number of minutes), lesson outcomes and 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 plans consist 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 provided the answers (but tells the teachers not to write these on the board.) The ten 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 (which is your favourite?); a few required slightly higher order engagement (why do they add sugar over 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 component, the "just-in-time" training at the start of the programme and the ongoing in-class coaching is viewed by the programme designers as pivotal in shifting 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, 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.

2.2 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 link 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 reasons. The responsibility or burden shifts to the teacher to keep up with the pre-specified timeframes. 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 to increase both the amount of time on learning tasks and intensify work on the tasks, thus allows for increased opportunities-to-learn and curriculum coverage. 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 allow 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 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' words 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. Not only did teachers experience how the learning tasks embedded in each lesson built on each other, but 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, resources and classroom management.

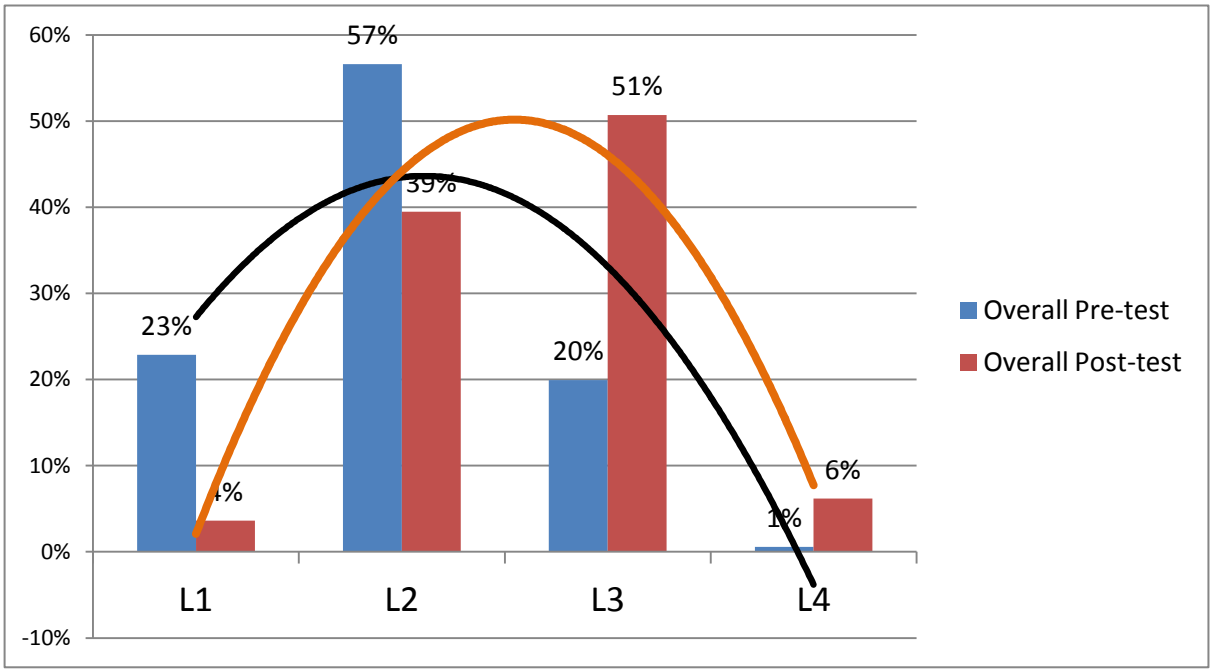
The third way it disrupts and re-engineers practice is that it links instruction more directly to the reading levels of most of the learners in the class. An emerging finding in international literature on large-scale reform is the negative consequences of the overambitious curriculum (Pritchett, 2012). 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.

2.3 Results of the *2012* Catch-Up Programme Pre- and Post-Test Study

A preliminary "pre and post" test evaluation suggested that the programme is effective (Hellman, 2012). That study was conducted internally by those responsible for the programme as it was administered in the Gauteng province. It focused on learner performance, assessing the extent to which the Catch-Up Programme improved four distinct literacy skills, i.e. 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 non-randomly selected sample consisted of 1570 classes, which was about 45% of English teachers covered by the programme. Hellman (2012) found that while not all learners were on the same level of achievement at the start of the intervention, across skills, NGOs and districts, the magnitude of gains made by learners was roughly of the same order. Overall, the programme's striking characteristic is that irrespective of grade, NGO and district, it seemed to have had a strong, positive and consistent effect.

Figure 1 is taken from Hellman (2012) and demonstrates the test score gains made by the children exposed to the Catch Up programme over the period of the programme. On the pre-test only about 21% of children scored above 50%, whereas after the programme about 57% of children scored higher than 50% on the test. The study had no control group with which to estimate what learning gains would have been in the absence of the programme.

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



Source: Hellman (2012)

2.4 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 an educational theory 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 compare findings. One of the problems with some of the existing South African studies is that the evaluations have often been undertaken by the programme developers, potentially compromising the independence of the investigations.

2.5 Sampling frame and rationale

The Pinetown District of KwaZulu-Natal province was the research site for the study. It has the advantage of containing a range of poor schools of different types (rural, urban, informal settlements and formal settlements). The study was conducted amongst grade 4 children in this district in schools where the dominant home language was not English. In the majority of cases the home language of children was isiZulu.

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 onward. The second criterion is that only schools that scored at 55% or below on the Grade 4 First Additional Language (FAL) test in both 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 on 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 control schools. In other words, all learners in a particular grade in a selected school were included in the study. The ethical reason 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, others will not. The practical reason was that if the study had a sub-sample for treatment or

control within a school, the language teacher would have 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 spill-over effect from the treatment to the control schools.

The study team made the following assumptions when planning the sample:

1. Each school is regarded as an intact cluster for the purposes of calculating standard errors.
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 2013 ANA) are included.
4. Only public ordinary schools are included.
5. 80% power level and 5% significance levels[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 pre-tests and post-tests of 0.7.
10. Attrition amongst 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 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 Pre-Test Learner Results

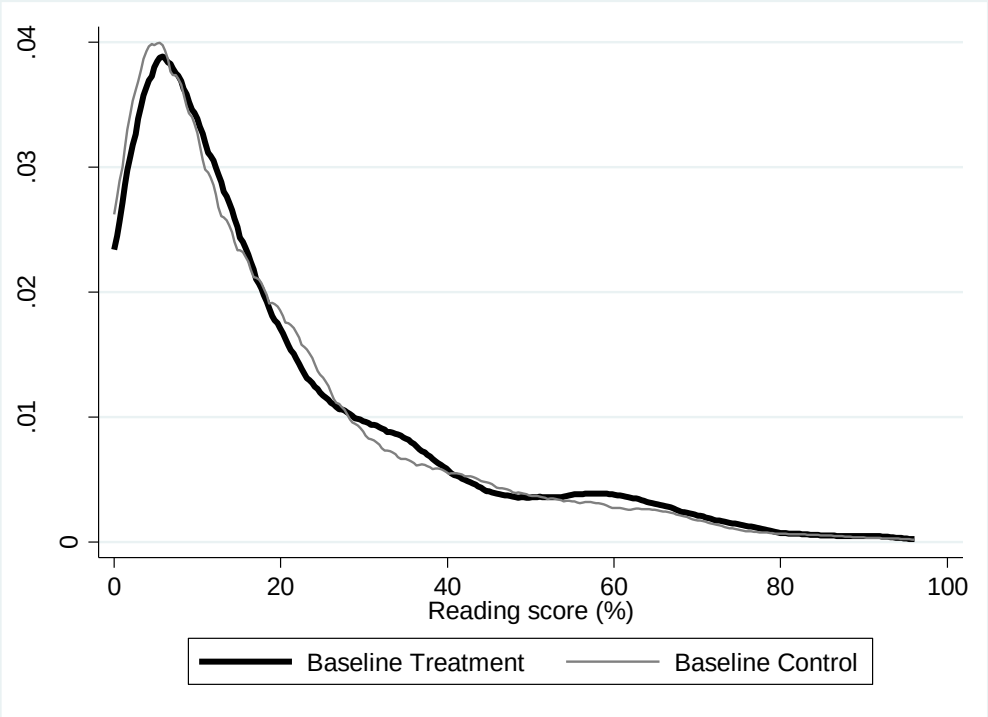
This sub-section begins with descriptive information on the intervention in Pinetown. This is followed by the presentation of key data from the pre-test.

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 pre-testing 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 pre-test for 2663 learners from 96 schools. For purposes of analysis, however, we only used data from the 2543 learners who also wrote the post-test. The focus of the data analysis of the pre-test was on the effectiveness of test items and to check the balance between the treatment and control schools.

There were 36 numbered test items, a few items with multiple components. As such the total test score was out of 51. The first analysis was designed to ascertain the number of learners with non-responses on items. Non-response could have been due to no answer provided or more than one response provided. 75% 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 both learners in 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 pre-test. Given the very low scores on the pre-test, 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 Pre-Test Scores, Percentage



The questions on the cover of the test instrument allowed the research team to analyze 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 Grade 4 Learners

Age	**Mean reading score**	**Number of learners**
8	27.09	11
9	21.80	1072
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
Total	18.41	2543

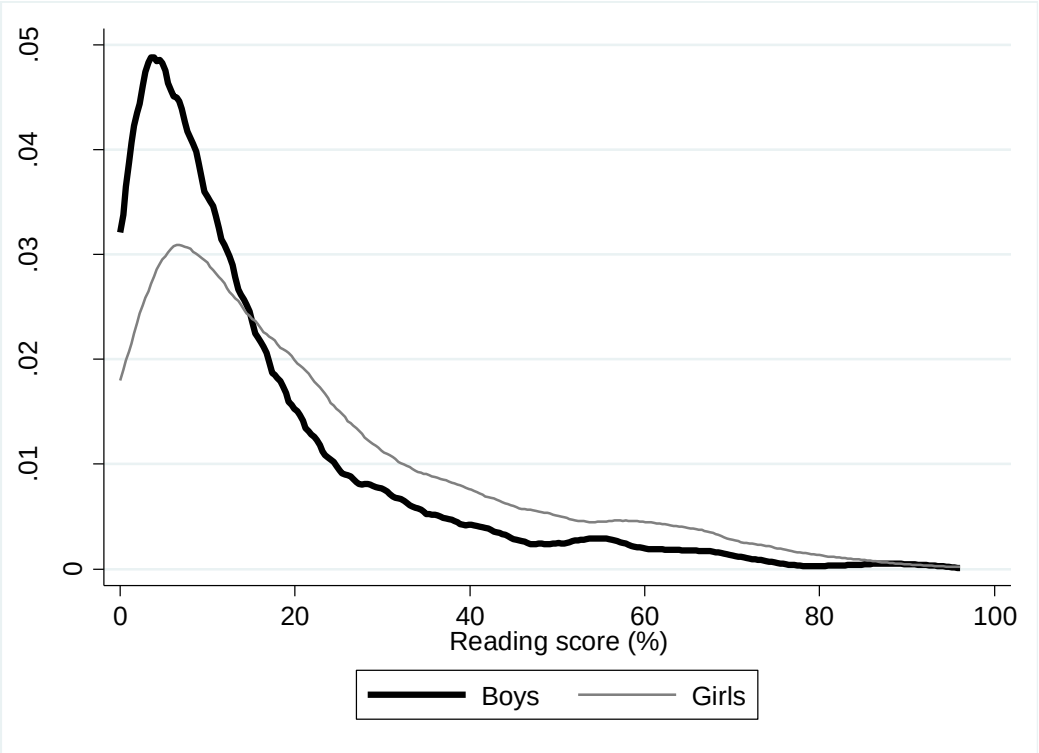
Table 1 reveals that on average, schoolchildren at the 'correct' age to grade had the highest mean scores, with the scores dropping substantially for older learners. What is of concern is the relatively large number of learners (148 out of 2543) who reported their age as 14 years or older, five full years beyond the norm for the Grade 4.

Table 2 Baseline Performance by Gender

Gender	Mean reading score
Boys	14.89
Girls	22.05
Total	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 South African test results, such as PIRLS 2011 and the Annual National Assessments of recent years, which all show a significant test score advantage for girls especially in literacy.

Figure 3 Distribution of Baseline Performance by Gender



2.7 Implementation

The service provider collected information about the enactment of the programme by teachers or, put differently, the levels of compliance with the programme. Altogether, 79 lesson plans could have been implemented over the period of the intervention, and on average, teachers completed 66% of these lessons. Five teachers completed fewer than half the lessons. Eight teachers completed at least 75% of lessons. We also have information on the number of class assessments (that were provided as part of the intervention programme) that each teacher completed. 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 3 indicates, some teachers only attended one or two afternoon workshops while others attended five or six.

Table 3 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 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 to some extent, but teachers who 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 post-testing would be implemented.
- Teachers needed support with the technical process of working out average test scores for reporting purposes. In response to this, the implementing agency introduced a 'reward system'. Once teachers were up to date with submissions, and the submissions have been verified against learners' books, they 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 regards to written work.
- The management of classroom resources by teachers 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.

3. Results

3.1 Attrition

From the perspective of the study design, one of the most positive outcomes of the post-test was the low level of attrition between the pre-test and post-test. No entire schools were lost on follow up. Table 4 shows that attrition amongst learners appears to have been low and not particularly skewed across treatment and control groups.

Table 4 Attrition between Pre-test and post-test, RCUP 2014

	Present at Endline	Not present at Endline	Total
Control	1423	127	1550
	(91.81%)	(8.19%)	(100%)
Treatment	1043	70	1113
	(93.71%)	(6.29%)	(100%)
Total	2466	197	2663
	(92.6%)	(7.4%)	(100%)

Overall, of the 2663 learners who wrote the pre-test, 2466 completed the post-test, which represents a 7.4% attrition rate. The attrition rate was slightly higher in the control group compared to the treatment group. When running a regression to test whether allocation to treatment group predicts attrition it is evident that Treatment does not predict attrition at all once controlling for variables such as baseline scores. Therefore, we delete learners that were absent from the dataset and proceed to analyse the data using only learners present in both the pre-test and post-test.

3.2 Main results

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 post-test means between control and treatment school groups.[8] A comparison of the trend lines in the pre and post tests for the treatment and control schools, shows that while both groups improved substantially between the pre and post tests, the improvement is only marginally better in the treatment group. In other words, while the base-line trends were very similar, so were the end-line trends.

Figure 4 Post-test Score Distributions for Treatment and Control Schools

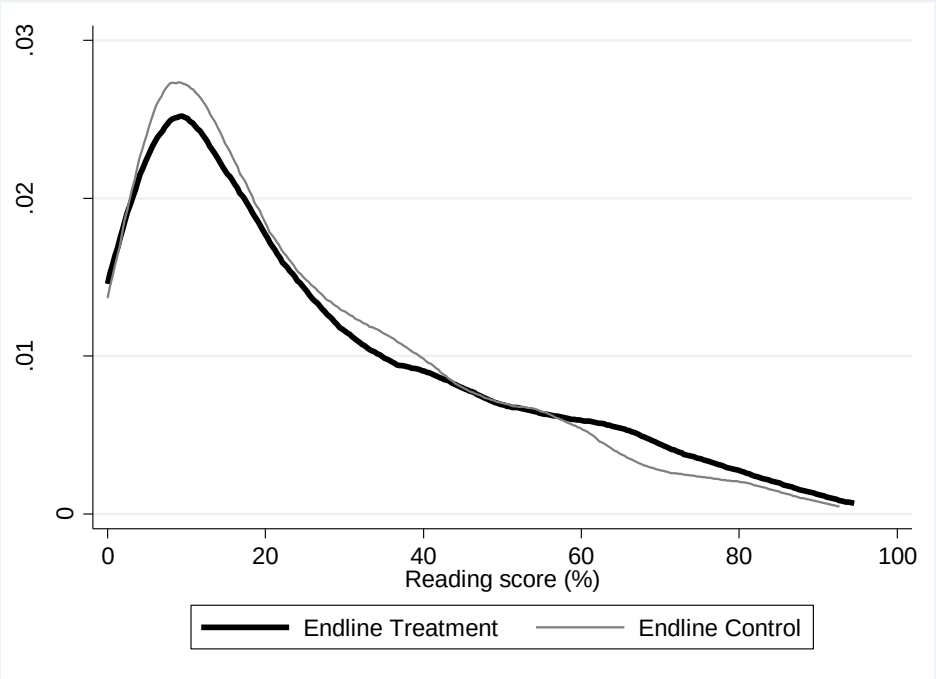
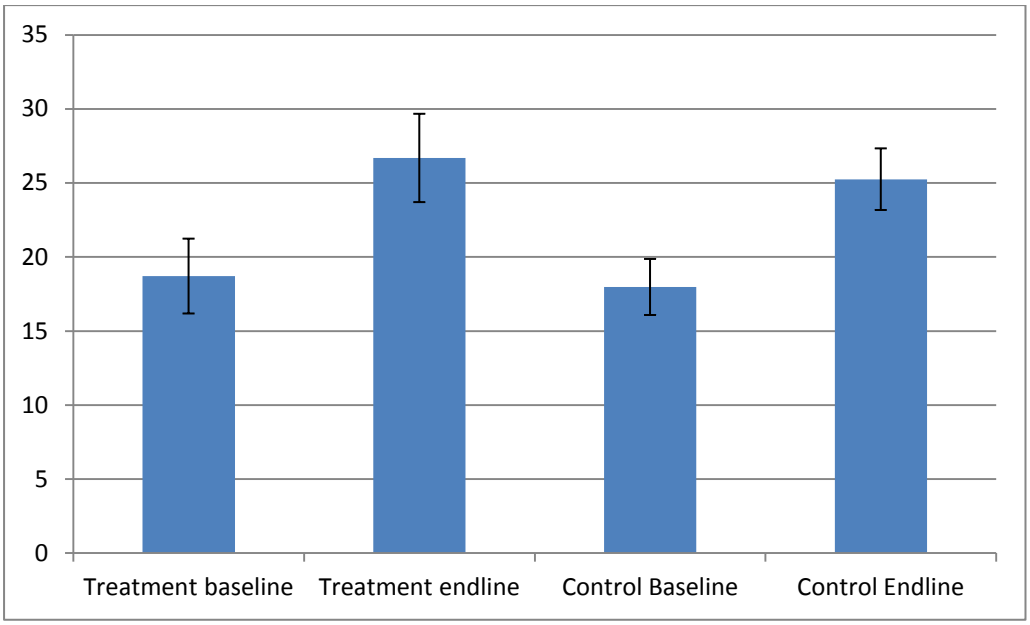


Figure 5 Mean Scores for Treatment and Control Groups (Pre and Post-test)



Note: 95% Confidence Intervals are indicated

The relative small difference between the improvement in the treatment and control schools is clearly evident in figure 5. In statistical terms, although the treatment schools mean post test score was higher than the control group, the difference is not statistically significant.

Table 5 shows the results of five regression models, which represent the most robust methods for estimating the impact of the programme. Column 1 shows the model where the outcome variable is the overall score on the post-test or end-line 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 pre-test score, stratification dummies, learner gender, 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 endline test scores to be different between treatment and control schools *other than because of 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 5, but all the above-mentioned controls were included. Columns (2)-(5) on the right of the table refer to models with the same set of explanatory variables but the outcome variables are learner scores for each of the four literacy domains which formed part of the reading test.

Table 5 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
SE	(0.67)	(0.61)	(1.07)	(1.34)	(1.40)
Observations	2466	2466	2466	2466	2466
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, 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 an additional 0.49 percentage points relative to the control group. However, we are unable to conclude with any level of statistical confidence that the true effect is different from zero. 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 different from zero.

3.3 Heterogeneous treatment effects

We also investigated so-called 'heterogeneous effects' – 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 not reported here). In planned forthcoming analysis we will continue to

investigate heterogeneous effects according to other characteristics as outlined in the Pre-Analysis Plan.

The following analysis (table 6), however, points to the possibility that the impact was larger for children who initially performed better on the baseline test. The result is statistically significant for spelling. Although not significant in language the size of the coefficient is actually larger than that for spelling so it may be that the same was true for language and we are simply unable to conclude so with statistical confidence. For spelling, there was effectively no impact on those who had initially scored poorly (and there were indeed many zero scores). The coefficient on the interaction term indicates that every additional 10 percentage points on the baseline test was associated with an increased treatment effect of 0.5 percentage points.

Table 6 Impact by Baseline Performance of Learners

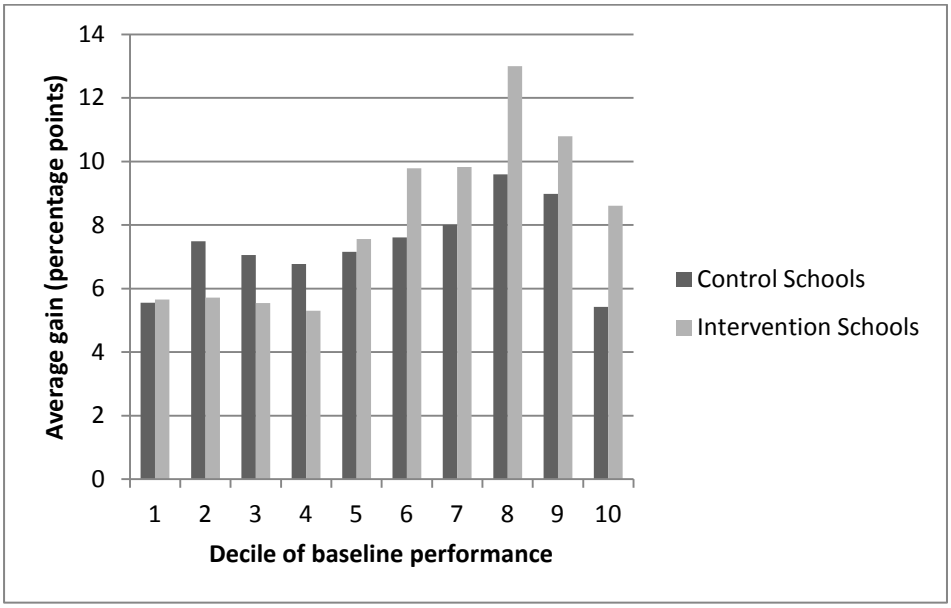
	Combined score	Spelling	Language	Comprehension	Writing
Treatment	-0.44	0.32	2.92**	-1.96	1.39
	(0.86)	(0.7)	(1.15)	(1.54)	(1.47)
Baseline percentage score	0.97	0.93	0.66	0.79	0.49
	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Treatment **X** Baseline	0.05	0.05*	0.07	0.02	-0.03
	(0.04)	(0.03)	(0.04)	(0.04)	(0.06)
N	2466	2466	2466	2466	2466
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, 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.

Although the regression analysis above did not conclusively indicate that programme impact varied according to baseline learner performance on the combined test score, some descriptive analysis points to the strong possibility that it did. The result in Table 4 may be a functional form issue. The following graph shows the average gain score for learners in treatment and control schools by each decile of baseline performance. Deciles are ten equal sized groups of learners split according to baseline performance. So, Decile 1 includes the bottom 10% of learners on baseline performance.

Figure 6 Average gain scores by decile of baseline test performance



3.4 Effects based on differing treatment intensity

The main estimate of the programme impact as reported in Table 5 is conventionally referred to as the "Intent 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), i.e. the effect of the intervention for those who complied with the intervention. In our particular situation we are not able to retrieve the TTE since compliance is not a zero or 1 categorization but rather there were varying levels of compliance. Therefore, we are only able to show descriptive statistics of the average learner gains depending on how many training sessions teachers attended (Table 7). Note that zero category includes the control schools. The gains were highest when teachers attended at least three training sessions, pointing to the possibility that the success of an intervention such as RCUP may depend on the extent to which teachers engage with it.

Table 7 Average learner gains by number of training sessions attended by the teacher

Number of sessions	Language	Spelling	No learners
0	5.37	4.63	1606
1	1.22	2.45	74
2	4.23	4.40	189
3	8.58	6.46	254

4	8.87	6.18	142
5	6.73	2.95	101

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 8 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. Firstly, the coaches were not randomly assigned to schools. However, the fact that we have baseline scores for each learners and can control for stratification and other learner, school and teacher characteristics reduces the likelihood of omitted variables bias. Secondly, the effective sample size is cut in half – instead of 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 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 8 Impact of 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	2466	2466	2466
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, 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.

3.5 Impact on Annual National Assessments

The Annual National Assessments (ANA) of 2014 were written during the week of 16 – 19 September across South African schools. This was about three months after the RCUP intervention was

finished. 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 as First Additional Language or Afrikaans as 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 spill over 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.[9] 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.

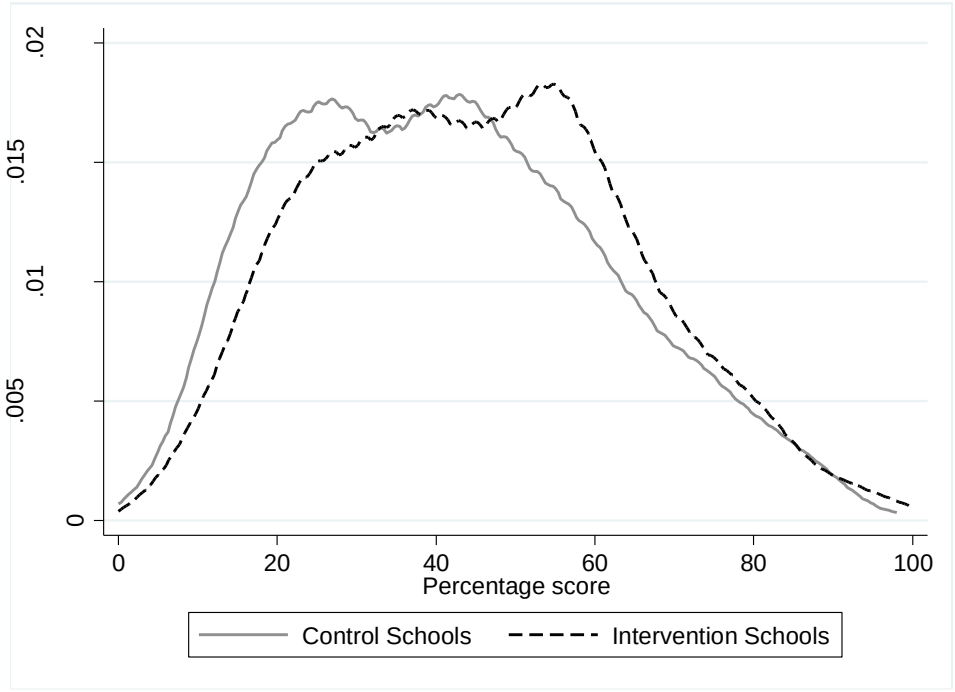
There are, however, several limitations of the ANA data for our purposes. The data quality is not expected to be as high as that collected by our service provider. This is because the ANA tests were locally administrated and marked by teachers within each school. Differences in the conditions of testing 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 end-line test score and the ANA language scores of learners. In a sample of 1928 learners who we were able to match between the RCUP and ANA datasets, the correlation coefficient between the baseline test score and the end-line 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 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 zero). Fortunately though, there is no reason to expect differences in marking or the 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, i.e. not only those learners who were sampled for our own independent testing. This provides us with a dataset of 6419 learners across our treatment and control schools. While this improves the statistical power for identifying 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 as a First Additional Language within our sample of schools was 43.0%. As was the case in our independently administered tests, girls (average score of 46.8%) substantially outperformed 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 (boys are noticeably older than girls on average). Although this finding is not central to this paper, it confirms an increasingly clear pattern of a large learning disadvantage for males in South African schools.

Figure 7 presents Kernel Density curves showing the distributions of test scores for those in intervention schools and those in control schools. This indicates that learners in intervention schools had a somewhat better distribution of achievement than those in control schools. This is a preliminary indication of a positive treatment effect.

Figure 7 Kernel Density Curves of Test Scores for Grade 4 English as 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 those 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.[10] Models 2, 3, 4 and 5 in Table 9 show the estimated treatment effect when different ways of controlling for prior school performance are used (controlling for school mean language score in ANA 2013, controlling for school mean language score in ANA 2012, controlling for both school mean language score in ANA 2012 and in 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 9 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 LANG 2013	No	Yes	Yes	No	Yes
School mean grade 4 LANG 2012	No	No	Yes	No	No
School mean RCUP baseline	No	No	No	Yes	No
School mean grade 3 LANG 2013	No	No	No	No	Yes
N	6419	6419	6419	6419	6055
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 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 and 54 control). 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 who had been exposed to the catch-up programme? Since the mathematics test is written 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 10, 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 10 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	6687
r2	0.2153

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

All models include controls for stratification dummies, learner gender and 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 11 indicate that there was a positive effect for the grades either side of the treated group, i.e. 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 at Grade 4. Therefore, we recommend that no strong conclusions be made on the basis of this result.

Table 4 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	9144	7673	7089	5341	4963
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 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 and 54 control). For grades 1, 2 and 3 the test was a home language test whereas 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, whereas 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 12 Impact of Coaches (ANA)

	ANA language
Coach A	5.38**
	(2.45)
Coach B	0.85
	(1.70)
N	6419
r2	0.2106

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

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

3.6 Analysis of sub-sample of individuals participating in both RCUP testing and ANA

Out of the 2466 learners with valid pre and post-test scores in the final RCUP dataset we were able to identify 1928 learners in the Universal ANA dataset of 2014. We matched learners 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 individuals were identical based on these variables. We therefore dropped all such individuals 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 present 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, i.e. they may have misspelt their name or surname 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 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 as the main model in Table 6 (i.e. predicting RCUP endline scores) but on the sub-sample of 1928 matched learners we obtain essentially the same estimated treatment effect (a coefficient of 0.48 as opposed to 0.49).

Table 13 reports the results of the two models 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 earlier results throughout 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 coefficients of interest in these two models are not statistically significantly different from zero. The effect sizes are non-negligible which means we were somewhat underpowered, especially 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 5 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	1928	1928
r2	0.4643	0.4676

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 the baseline score of the learner in the RCUP testing, stratification dummies, learner gender, 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. Discussion

Even though the increases in the learner 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 standardized scores were relatively small compared to the gains suggested in the original 2012 Reading Catch-up study and in Pretorius' (2014) new study[11]. A scan of a sample of learners' post-test scripts from amongst 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, comprehension and writing. 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 'sub-system' 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 like 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 around domains such as spelling and language. The potential for improvement through this sort of programme appears larger for those learners who are not at the very bottom of the performance distribution.

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 pre-test and post test, the gains was 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. Firstly, it may simply be that soon after beginning with English as the language of instruction (as occurs in grade 4) learners typically demonstrate quick 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 or the treatment group, they were all subject to external scrutiny particularly around learner performance testing (i.e. pre and post testing). 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 through having been 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 spill-over 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 were not easily transferable (coaches and materials).[12]

A third explanation may be found in the 'floor-effect' evident in the pre-test 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 counterparts in Gauteng. A different instrument, one that emphasized 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, 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 words directly during the ten weeks of lessons or mastered some aspect of English phonics, it would be reasonable to expect that this learning would be evident in the post-test and a few months later in the ANA test. Similarly, explicit teaching of basic language structures, such as capital letters at the beginning of the sentence and full stop at the end, would carry through to improved scores on the language section of the post-test. In contrast, the fact that comprehension scores did not change, 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 a often hidden but powerful factor.

Notwithstanding this strong finding, the study has also provided substantial evidence around a range of themes. These included further evidence of the serious under-performance in English as a first additional language at the start of the Intermediate Phase and the scale of the gender performance gap.

The study pre-test dataset suggested that the Grade 4 learners' English language knowledge and skills 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 the performance levels indicated by 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 administrated and marked.

Another major insight from the pre-test analysis is the large performance gap between boys and girls. This gap is evident both in the pre and post tests and is consistent between 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 pre- and post-test 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, that improved performance cannot be simply ascribed to the intervention. The study also shows the value of replication studies to address questions of external validity. Assuming the results of the Gauteng study were reliable and valid, this study demonstrates 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 fertile ground 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 first additional language literacy acquisition in our schools.

References

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

Cohen, D. K. (2011). *Teaching and its predicaments*. 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 Project1. *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*. Routledge.

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

Hopkins, D. (2003). *School improvement for real*. Routledge.

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*,

National Education Evaluation and Development Unit (2013)

Perry, H., & Fleisch, B. (2006). Gender and educational achievement in South Africa. In V. Reddy (Ed.), *Marking matric: Colloquium proceedings* (pp. 107-26). 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. (2015). *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*. Dorchrecht, Netherlands: Springer

Taylor, N, van der Berg, S & Mabogoane T (eds.) (2013) *Creating Effective School.* 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.

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 toward the measurement of certain literacy skills rather than oral reading fluency and comprehension.

[2] The terminology of "treatment" and "control" groups 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 "intervention" group and "treatment" group interchangeably.

[3] A paper by Kremer, Brannen and Glennerster (2013) provides a concise review of international RCT studies focussing on education.

[4] Initially, we tried to select schools based on the original ANA 50% and 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 (i.e. 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 with-school variation amongst students. 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 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 is has been standardized to a common effect size unit, i.e. 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] An internal DBE analysis of the Annual Survey of Schools indicated this.

[10] 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. Moreover, we include the strata dummies in the regression to further control for differences in school characteristics, including prior ANA achievement, which was one of the dimensions influencing stratification.

[11] Although Pretorius' study used only one school and therefore should not be considered as a benchmark for typical effect sizes.

[12] Lemon (2014) describes 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.