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:

$$DID = (\bar{y}_{T,t=2} - \bar{y}_{T,t=1}) - (\bar{y}_{C,t=2} - \bar{y}_{C,t=1})$$

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 90th 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 10th 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).