Art. # 1248, 13 pages, doi: 10.15700/saje.v37n1a1248

The value of large-scale randomised control trials in system-wide improvement: The case of the Reading Catch-Up Programme

Brahm Fleisch
Division of Educational Leadership, Policy and Skills, Wits School of Education, University of the Witwatersrand, Johannesburg, South Africa
brahm.fleisch@wits.ac.za
Stephen Taylor
Research Directorate, Department of Basic Education, Pretoria, South Africa
Volker Schöer
African Micro-Economic Research Unit, School of Economic and Business Sciences, University of the Witwatersrand, Johannesburg, South Africa
Thabo Mabogoane
Department of Planning, Monitoring and Evaluation, Presidency, Pretoria, South Africa

This article illustrates the value of large-scale impact evaluations with counterfactual components. It begins by exploring the limitations of small-scale impact studies, which do not allow reliable inference to a wider population or which do not use valid comparison groups. The paper then describes the design features of a recent large-scale randomised control trial (RCT) evaluation of an intermediate phase literacy intervention that we evaluated. Using a rigorous sampling process and randomised assignment, the paper shows the value of the approach, and how the RCT method prevents researchers from reaching potentially harmful false positive findings. The paper also considers some of the limitations of the RCT method and makes recommendations to mitigate these.

Keywords: impact evaluation, intermediate phase literacy; Randomised Control Trial

Introduction

The South African education system has achieved significant success in ensuring almost universal access to basic education, but improvement of poor reading skills and mathematics performance remain elusive goals. In response, government has introduced and implemented a variety of educational policy changes at national and provincial level. Unfortunately, little is known about the impact that these policy changes have on learning outcomes. With limited resources available to us, it is crucial that we understand what research tells us about which policy changes work and which do not. Specifically, the policy-maker needs to know the causal impact of particular programmes and policies on the intended outcome of interest. While it is valuable to obtain a detailed understanding of how, when and why policies work in very specific contexts, it is of primary importance for the policy maker to know the average impact on the population as a whole, and perhaps also the average impact on relevant sub-groups.

Although the number is still small, there are a growing number of rigorously designed quantitative education impact studies being conducted in South Africa (Mouton, 2013). While many of these studies are responses to funders' mandated project or programme evaluations, a number of recent studies have been initiated for the purpose of advancing knowledge of how to improve instructional practices and, by extension, how to improve education outcomes in South African schools. But for this new body of scholarship to gain traction, it is imperative that it meets certain research standards. Specifically, impact evaluations need to be based on a sufficiently large and representative sample from the targeted population to ensure that we can be confident that the findings of our study tell us something about the policy intervention and the target population. In order to measure the causal impact of programmes or policies, these studies also need to include a valid estimate of the 'counterfactual' in order to avoid false positive and false negative findings.

The purpose of this article is to contribute to this emerging trend in South African education research by illustrating the value of large-scale impact evaluations with counterfactual components. The second section of the article begins with an analysis of two published studies of intermediate phase (Grades Four to Six) literacy interventions. While these studies provide important research insights, their designs demonstrate the limitations inherent in small-scale studies that use either no comparison group or a convenience comparison group as a counterfactual. In the third section, we describe in some detail the design characteristics of a recent large-scale randomised control trial evaluation of an intermediate phase literacy intervention that we evaluated. Using a rigorous sampling process and randomised assignment, we show the value of the approach and how the RCT method prevents researchers from reaching potentially harmful false positive findings. We are aware of the shortcomings of the randomised control trial as a research method. Building on the work of Peters, Langbein and Roberts (2015), the fourth section explicitly discusses the weaknesses of RCT designs and identifies two new challenges – what we call the leadership effect, and the spill-over paradox.[i]

Studies from the South African Literature

How do South African researchers establish whether or not the intervention which they investigate affected the learning experience of learners, i.e. that the intervention had an impact? To answer this question, we analysed two examples of published evaluation studies that focus on the subject and school phase of this study, remediating intermediate phase (Grades Four to Seven) literacy. While there are other, similar studies, these examples have been selected in order to illustrate methodological aspects of research in the field in South Africa. Both focus on interventions aimed at improvement of a specific element of literacy, i.e. comprehension skills, for disadvantaged learners in the intermediate phase.

Klapwijk and Van der Walt (2011) evaluated the impact of an intervention designed to show teachers how to use a comprehension instructional framework that focused on 'starter' comprehension strategies. The intervention itself consisted of a single session of training, where various resources were provided, such as a teacher checklist, a booklet, additional information on intervention strategies, lesson samples, handouts, story maps, and summary sheets. The intervention also included ongoing support as needed from the implementer/researcher. The intervention took place over 15 weeks, starting with the administration of a pre-implementation test and ending with a custom-constructed post-intervention test. Of particular interest was the study's sampling frame. The study of the efficacy of the comprehension instructional framework was undertaken in two Grade Five classes of the same school. The researchers described this school as serving low socio-economic status learners in a predominantly Afrikaans-speaking community. Although the school was predominantly Afrikaans, there were two English Home Language Grade Five classes with different teachers in the year that the study was conducted. The researchers selected one of the classes for the intervention and the other was designated a 'control' group. The researchers maintained that assigning the two classes in this way made it "possible to use a control group and gather data within the same grade for comparative purposes both before and after the intervention" (p. 29). The only contextual data provided on the learners in the two groups was the number of boys and girls in each group. The tables in the results section present the differences between the treatment/experimental classroom and the control classroom pre- and posttest results. The final "effect-size analysis" table presented disaggregated results on the posttest, comparing the intervention classroom learners' gains to those in the control classroom. While the authors do not make any explicit statistical or policy claims about the 'effectiveness' of the intervention, they do imply that they have "evidence that it is possible to measure the transfer of strategy knowledge" and that this evidence would motivate teachers to adopt the intervention.

The second article also reports on an intervention designed to improve intermediate phase literacy. As part of a larger research study, Pretorius and Lephalala (2011) evaluated an intervention that was designed to assist Grade Six teachers teach comprehension of narrative texts more effectively. In this instance, the study was not a control/treatment design, but rather a comparison design in two schools. The Treatment 1 school had a voluntary after-school programme intervention. The Treatment 2 school was provided with a comprehension programme that was implemented during the formal school timetable. Pretorius and Lephalala (2011) recognised that the treatment groups, while they shared similar characteristics, such as school size, lack of resources, poor infrastructure, and print-poor classroom environments, were different in two critical respects, namely: their quintile status and the language policy. In the results section, the authors presented the tabular results which differentiate the performance on pre- and posttests in the two 'groups'/schools. Based on the results from the in-class teaching school (described as the intervention group in Table 1), the authors state that the intervention undertaken during the formal school timetable led to substantial gains (with a large effect size) in English comprehension.

In terms of research design, these two studies share in common the following design limitations:

1. both have an inadequate description of the study sampling frame and the logic that links the study sample to the wider school population;
2. the schools were likely to be chosen for convenience, which renders them less likely to be representative of the target population of schools; and
3. the comparison groups are systematically different to the treatment groups in a number of aspects such as demographic composition of learners, administration, and teacher characteristics. They are therefore less suitable as a means of comparison.

When offering a critique of these studies, however, we need to be cautious not to conflate policy effectiveness, methodological rigour, sample equivalence, and analytic approach. Clearly these studies were not intended to offer definitive evidence that could be taken as a 'policy warrant', something that would require a higher level of methodological rigour. While not made explicit, the studies are part of ongoing research programme that extends prototype pilot intervention pre-and posttest studies with 'rough' or illustrative equivalent control groups.[ii] In the next section, we draw the distinction between consequential and illustrative counterfactuals to show the potential value of both.

On the question of analytic approach, a meaningful distinction can be drawn between studies that use features of experimental design but are not intended to produce policy evidence, and more robust approaches that are explicitly designed for this purpose. That said, we clearly recognise that even with robustly designed experimental studies, evidence of policy applicability require further studies of implementation feasibility.

Pilot studies, pre-posttest designs, quasi-experimental, and qualitative studies, are crucial in identifying problem areas in the first place, field testing prototypes and allowing researchers to develop hypotheses and to discover generative mechanisms of change. If regarded as 'pilot' or prototype studies of interventions, the studies described above can demonstrate the feasibility of the intervention concept and highlight implementation issues. These small-scale studies might also provide insight into teachers' perspectives of the interventions and provide insights about typical practice.

Because of their study design, we ought to be cautious in attributing any change in the learners' literacy performance to the learners' exposure to the intervention. To claim effectiveness in the absence of adequate sample sizes and valid control groups could lead to potentially wasteful or even harmful decisions. Attribution of changes in performance to a particular intervention would require a rigorous impact evaluation design. We therefore argue that both qualitative and quantitative research methods are necessary parts of the entire research agenda, as they complement one another throughout the research process, each with its own advantages and disadvantages.

The remainder of the article uses the case of the Reading Catch-Up Programme (RCUP) to illustrate the challenges, complexities and ultimately the values of rigorously designed randomised trials, i.e. studies that have three key features:

1. sufficiently large sample size for hypothesis testing;
2. a randomly drawn sample from the target population to ensure that the sample is representative of the target population; and
3. randomised assignment to the intervention for a valid counterfactual.

While we believe that the RCT has real value, we are certainly not oblivious to the weaknesses or limitations of these types of studies.

The Evaluation Problem and the Search for the Counterfactual: Illustrative v. Consequential?

Every study that aims to investigate the impact of an intervention faces the same evaluation problem: what would have been the outcome if the schools or learners had not been exposed to the intervention? Would they have performed differently? If so, by how much? By how much did the intervention change the learners' performances compared to how they would have performed if they had not been exposed to the intervention?

To estimate the impact of a programme or a policy with some certainty we would need to measure this counterfactual scenario. Unfortunately, in the real world, once someone has been exposed to the intervention, it is impossible to observe the counterfactual. We simply cannot turn back time, to do it all over again, but without the intervention. The best we can do is construct an estimate of the counterfactual situation, i.e., we need to use the outcome of a comparison group that we consider to be as close as possible to the outcome the treated group would have had if not exposed to the intervention. The method we use to do this will determine how convincing the estimate of the counterfactual will be. Let us use extra mathematics lessons as an example of a programme whose impact we are interested in knowing. One might use a before-after analysis like Hellman (2012) and compare test scores of learners who attended extra lessons before (time 1) and after they had attended the programme (time 2). However, it is to be expected that learners will improve their mathematics knowledge over time in any case: through their core school lessons and through a range of other influences, including becoming more mature. In this case, simply comparing pre- and post-scores tells us little about the impact of the programme – we would have needed to observe their scores (at time 2) had they not attended the lessons, which is obviously impossible.

Alternatively, one might compare learners who attend extra lessons to learners who do not attend extra lessons. However, these would most probably be two rather different groups of children. If those who attend extra lessons do so because they are performing poorly, and realise they need extra help, then we would expect these learners to perform worse than those not attending lessons. If we regarded learners not attending extra lessons as an estimate of the counterfactual, we might conclude that extra lessons had a negative impact. This would of course be a false conclusion, because it relied on a comparison group that is systematically different from the treated group, and therefore, would provide us with an invalid estimate of the counterfactual.

There are a number of quasi-experimental methods that attempt to construct a counterfactual in more sophisticated ways, and (depending on the situation), these will be more or less convincing. For example, one might compare pre- and post-outcomes for both programme beneficiaries and non-beneficiaries, thus effectively comparing the gains over the duration of the programme. This is known as the differences-in-differences method. But one has to assume that the rate of learning would have been the same between the two groups in the absence of the programme. If, for example, the type of learner who takes extra maths lessons typically learns at a different rate as learners who do not typically take extra lessons, then it will not be valid to regard the difference in gains as a reflection of the true impact of the programme.[iii]

So how do we find a suitable comparison group whose outcome we can use to establish a counterfactual? The simplest and most convincing way to construct an estimate of the counterfactual is to assign a group of individuals to an intervention (or 'treatment') group and a comparison (or 'control') group using a lottery. This 'random assignment' ensures that there is no reason to expect the treatment group to be systemically different from the control group. Since receiving the programme would be completely random, the two groups should be similar in all observable characteristics. But even more powerfully, if no individual could choose to be in either group, the two groups should also be similar in all unobservable characteristics. It is for these reasons that the RCT design is considered the most reliable way of constructing a counterfactual.

Once we have a valid control group, a second requirement is a large enough sample size to smooth over any chance differences that might occur. For example, suppose one used a lottery to allocate two schools to a treatment group and two schools to a control group. It is quite possible that one high-performing school with an inspiring principal in the sample would skew the comparability between the two groups. However, with 100 schools randomly selected for each group, it is very unlikely that any such factor would be systemically different across the two groups. There are statistical formulae, which calculate the required sample sizes for such experiments with strong predictability (in a subsequent section we expand on the factors that influence the required sample size when conducting a Randomised Control Trial).

Possibly the best-known educational study that used an RCT to estimate the counterfactual was the Finn and Achilles (1990) analysis of the Tennessee Student Teacher Achievement Ratio experiment (cited in Green, Camili & Elmore, 2012). In the experiment, learners were assigned to one of three groups. The first group consisted of regular class sizes (22–25 learners), which was the control group. The second was the main treatment group, with substantially reduced class sizes (13–17 learners). The third group was an alternative treatment, which had the regular class sizes, but also included teachers' aides. When the average scores of 80 schools in the study were compared, students in the reduced class size fared better than either the control group or the alternative treatment. The learners in the classes with the aides did no better than learners in the classes of regular size. The researchers were able to conclude – given the sample size and random assignment of schools to control and treatments – that the policy option of reducing class size in the Tennessee context caused academic achievement to increase, and was superior to the other policy option, which was providing a teacher's aide to classes of regular size.

We make a distinction between illustrative counterfactuals and consequential counterfactuals. While a number of studies include illustrative counterfactuals (Klapwijk & Van der Walt, 2011; Pretorius & Lephalala, 2011), their purpose is to show what a 'typical' school might be like compared to the intervention school. In contrast, a consequential counterfactual allows the research to measure the impact; to provide a more precise estimate of the impact that an intervention may have on learning outcomes.

In our view, two key criteria need to be met for consequential counterfactuals. First: study subjects, (whether schools or learners), are assigned to the treatment and control groups on a genuinely random basis. We have found that researchers and policy-makers often use the term 'random' carelessly to describe how a programme was allocated. What they really mean is that district officials allocated schools on some unknown basis or schools themselves chose to participate for reasons of their own. In fact, the strength of an RCT lies in the fact that the researcher knows exactly how programme assignment was done and can therefore be certain that the control group provides a valid comparison. In some special settings, a natural experiment occurs, and shows that assignment to a particular programme or resource was effectively the same, as if done through a lottery. In such cases, this first criterion of random assignment is met. However, convenience sampling does not meet this criterion.

The second criterion is that there must be an adequate sample size in order to estimate the programme impact within a narrow enough band of uncertainty. Given that the purpose of statistical inference is to be able to draw conclusions about the target population based on an estimate that is calculated for a representative sample of the target population only, we need a sample that is large enough for us to have confidence in the estimate. There are various factors that influence the required sample size, but, to simplify, the larger the sample the more precisely one can make inferences about the population. Furthermore, the sample itself should be drawn randomly from the target population in order to be representative of the target population. The same is true of an RCT – one determines the impact of a programme based on a certain outcome, for example test scores. With only two treatment and two control schools, we will not be certain about whether the observed difference between the two groups after the intervention is an accurate estimate of the average impact if the full population of schools received the programme. However, if 100 schools are in each group, it will be possible to predict the true impact of the programme within a very narrow range. Statistically speaking, the standard error is a measure of the dispersion of a particular estimate, and the dispersion reduces as the number of observations increases.[iv] Unfortunately, there is no fixed number that researchers can use, i.e., an adequate sample size is not a certain percentage of the population. Rather, the optimal sample size depends on the variability of the outcome variable in the population, the smallest effect that one wants to be able to identify, and the confidence level at which one wants to avoid type one (false positive conclusion) and type two (false negative conclusion) errors when we test our hypothesis.

The required sample size in education research is typically larger than in many other disciplines, because learners are clustered in schools. To illustrate, consider two different samples, each of 200 learners, in South Africa. The first sample was obtained by randomly selecting 200 learners from all the learners in South African schools. The second sample was obtained by randomly selecting two schools in South Africa and then surveying 100 learners in each of these schools. Of course, the second sample is easier to access, but will provide a much less reliable representation of the South African learner population. This is because learners within a school are typically a fairly homogeneous group. In practice, this means that standard errors must be adjusted when sampling is done at the school level. The number of clusters (schools) and the size of each cluster (number of learners sampled per school) will influence the required sample size in order to measure with a satisfactory degree of precision.

From our study of the Reading Catch-Up Programme, we illustrate that in the absence of a consequential counterfactual (as the one implemented in this study), flawed conclusions could have been inferred about the programme. The flawed conclusion would lead to incorrect policy decisions, which could have had adverse effects, particularly on the weakest learners in the study population.

The Case of the Reading Catch-Up Programme RCT

In 2012, the Gauteng Department of Education developed and implemented an Intermediate Phase Catch-up Programme that aimed to close the learning gaps between the minority of learners who were reading at curriculum level and the majority that were reading far below the level. The English Catch-up Programme contains three key elements: scripted lesson plans, high-quality graded readers, and training and instructional coaching. The daily lesson plans provided a comprehensive description of each lesson. Coaches provided training to teachers in small groups, and they visited classrooms to model teaching practice and to observe, support, and encourage teachers as they worked on the lesson plans. They also monitored and tracked compliance. Using a simple pre- and posttest design, an unpublished evaluation (Hellman, 2012) showed that learners who participated in the programme improved their test score from an average of 24 to 40 percent. These positive results were confirmed in Fleisch and Schöer’s (2014) quasi-experimental study. Their results also indicate a positive effect on literacy scores. However, the authors warn of a possible test instrument effect that might drive the positive results of the previously underperforming schools. While the results are generally promising, a range of questions remain unanswered. Given the limitations associated with a simple pre- and posttest design in Hellman’s study, questions were raised about the veracity of the evidence of the impact of the intervention. Also, if the intervention did have an educationally meaningful impact, could it be replicated in other contexts? To address these questions, the researchers initiated the Reading Catch-Up Programme (RCUP) study, which replicated the original intervention using a more robust design in a different context.

The work of the RCUP study was divided into three parts. The research team, which included the authors of this article, designed the RCUP study itself, analysed data and reported on the findings. An education NGO was contracted to implement the intervention in treatment schools.[v] A totally separate evaluation NGO was contracted to collect learner information from pre- and posttests in both treatment and control schools. The Pinetown district of KwaZulu-Natal Province was the research site for the study. It was appropriate in that it had a large number of poor schools of different types (rural, urban, informal, and formal). It was also conveniently located close to the urban hub of Durban. The funder was engaged in a larger intervention aimed at improving school primary language and mathematics in the district. The Reading Catch-Up Programme could therefore have provided useful evidence, and if the study had shown strong positive results, the programme would have been rolled out to comparable primary schools in the district.[vi]

Particular care was taken in designing the most appropriate sampling frame and sample size for the study, to ensure that we were able to draw a representative sample of schools from the target population, achieve sufficient statistical power to identify a minimum detectable effect size, as well as satisfy ethical and cost concerns. The intervention was aimed at a target population which had functional but underperforming primary schools that were likely to have the infrastructure in place to respond to the intervention. Therefore, because the primary criterion of the intervention was to remediate English reading achievement of under-performing primary learners, we selected only those primary schools where English was the language of learning and teaching (LOLT) from Grade Four onwards. The second criterion was that only schools that scored 55% or below on the Grade 4 First Additional Language (FAL) test in both the 2012 and 2013 Annual National Assessments (ANA) tests in the Pinetown district were eligible for inclusion in the study. The third criterion was that selected schools must have entered between 15 and 120 learners in the FAL Grade Four ANA test in 2013 (a few schools actually exceeded this number in 2014). This was justified on the grounds of cost. One of the two biggest cost drivers in this intervention were learner support materials (particularly the graded readers, the number of which is determined by learner numbers) and coaches' salaries. 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.[vii]

For ethical and practical reasons, we sampled all Grade Four classes in the treatment and control schools. In other words, all learners in the particular grade in a selected school were included in the study. The ethical reason for doing this was that sampling classrooms within schools would mean that some schoolchildren would receive the benefits of the treatment within a single school and grade, and others would not. The practical reason was that if the study had a sub-sample for the treatment or the control within a school, and if the school had a specialist language teacher, she would have had to teach two different methods simultaneously, which would substantially add to the workload. We also wanted to reduce spill-over effects, whereby learners in non-treated classes of the same school might try to become exposed to the intervention by sharing resources with friends in treated classes. We assumed that, given the size of the province and the relative isolation of many rural schools, there would be little danger of a spillover effect from the treatment to the control schools.

One of the vexing questions that the researchers grappled with was the number of schools required to ensure that the study could have adequate statistical power. In order to arrive at the required sample size, the study team made the following assumptions:

1. Each school would be regarded as a cluster.
2. There would be an 80% power level, and a 5% significance level.[viii]
3. Testing would be restricted to a random sample within a single grade.
4. There would be an Intra-Class Correlation co-efficient value (between-school variance as a proportion of total variance) of 0.20.[ix]
5. Oversampling of control schools relative to intervention schools would be done in order to gain statistical power but save intervention costs.[x]
6. A correlation would exist between pretests and posttests of 0.7.
7. Attrition among learners would not pose a problem to the integrity of the study. Since the pre- and posttesting occurs within a 12-week period, absenteeism was probably going to be the main cause of attrition, and this would not likely to be systemically different between treatment and control groups. Consequently, attrition would not bias the estimated treatment effect.
8. The minimum detectable effect size (MDE) was set at 0.2 of a standard deviation in test scores.[xi]

Using these assumptions, statistical formulae were applied to calculate that a sample size of 40 treatment schools and 60 control schools ought to be adequate. A computerised lottery was used to randomly allocate the 100 sampled schools into the treatment and the control groups. These sampling assumptions ultimately 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.

We obtained data on the pretest for 2,663 learners from 96 schools. For purposes of analysis, however, we only used data from the 2,543 learners who also wrote the posttest. The comparison of the means and distribution of pretest scores indicates that the treatment and control groups were almost identical, confirming that the randomisation was successful in generating two similar groups. It was also clear that the vast majority of learners in both groups scored extremely low on the test, confirming the existing literature on literacy achievement (Figure 1).

Findings

The core question that animated this study focused on the extent to which learners' achievement in English literacy improved as a result of exposure to the Reading Catch-Up Programme. An analysis of the pre- and posttest results in the treatment schools showed that the learners whose teachers used the RCUP programme scored dramatically higher on the posttest, albeit off a very low base. The average learner score increased dramatically (from 18.7% to 26.7%), a gain of nine percentage points, percentage gain of just over 40% (Table 1). Any programme that shows an aggregate effect of improving learner performance by 50% within ten weeks would certainly be worth scaling up.

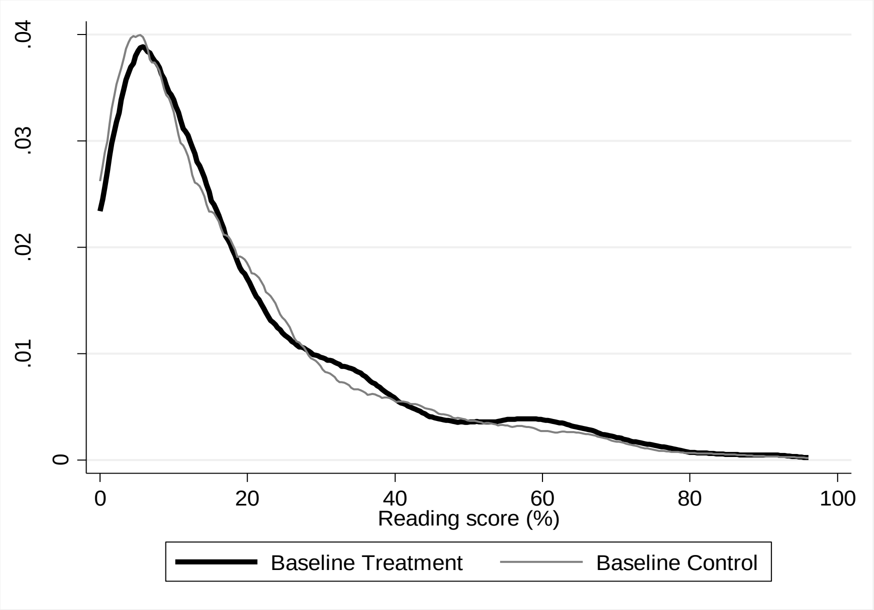

Figure 1 Kernel density of pretest scores, percentage

Table 1 Pre- and posttest scores in the treatment schools

	Treatment schools		
	N	Mean	St. dev.
Pretest	1,043	18.7%	18.5
Posttest	1,043	26.7%	22.6

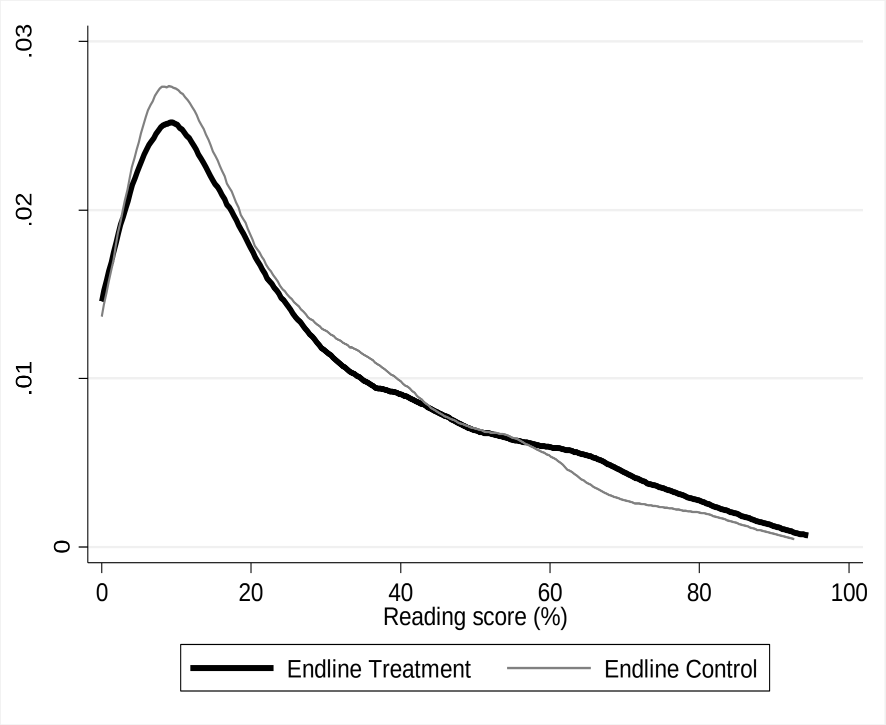

Figure 2 Posttest score distributions for treatment and control schools

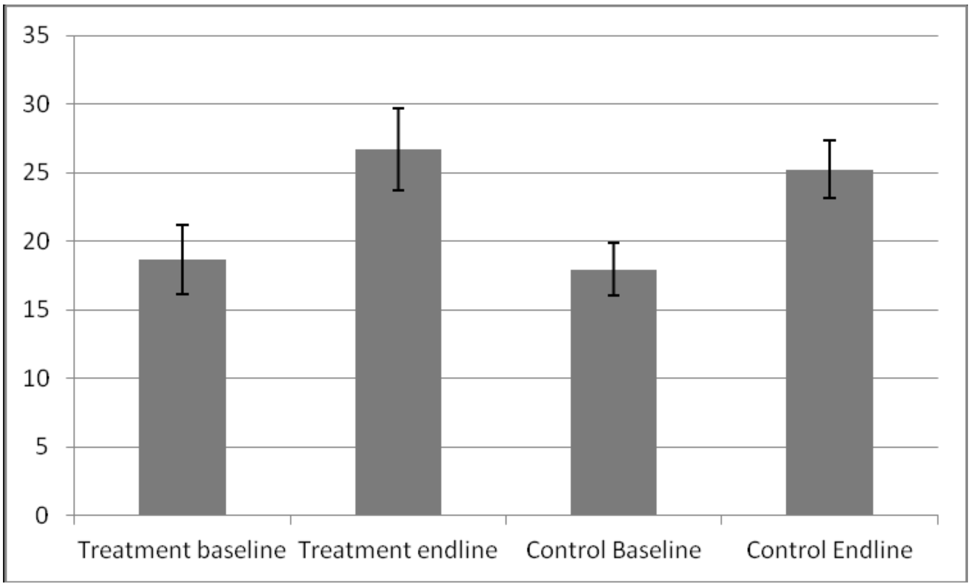

Figure 3 Mean scores for treatment and control groups (pre- and posttest)
Note. 95% Confidence Intervals are indicated.

However, when we add the counterfactual, i.e. the control schools, a fundamentally different picture emerged. The data showed (Figures 2 and 3) only a very small difference in posttest means between control and treatment school groups.[xii] A comparison of the trend lines in the pre- and posttest for the treatment and control schools showed that, while both groups improved substantially between the pretest and the posttest, the improvement was only nominally higher in the treatment group. In other words, while the baseline trends were very similar, so were the endline trends.

The small difference in improvement in the treatment schools relative to the control schools – the treatment's marginal impact – is visible in Figure 3. In statistical terms, although the posttest score was higher in the treatment schools than in the control group, the difference is not statistically significant. Thus, without the control group, one might have falsely concluded that the intervention made a difference. However, because schools were randomly assigned to the treatment and control groups we have a valid estimate of the counterfactual, namely the amount of learning that would have taken place over the period in the treatment group had they not received the intervention. This example illustrates the importance of a valid counterfactual estimate. Any number of RCTs could have been used to illustrate this. For example, Table 2 in Banerjee, Cole, Duflo and Linden (2007:1246) shows that both treatment and control groups in an education RCT conducted in India scored higher on the posttest than on the pretest on a range of test score outcomes. Banerjee et al. (2007) therefore compare the gains between the treatment and control groups in order to estimate the impact of the intervention.

Table 2 shows the results of five regression models, which represent the most robust methods for estimating the marginal impact of the RCUP programme. Column 1 represents the model where the outcome variable is the overall score on the posttest or endline literacy test. The main explanatory variable of interest is a variable indicating whether the school is a treatment school or a control school. Other variables included in the regression model are the learner's baseline or pretest score, stratification dummies,[xiii] learner gender, learner age, exposure to English at home, frequency of having an adult read at home, class size, teacher age, teacher gender, teacher qualifications, and school size.[xiv] Only the coefficient on the treatment variable and the standard error of the estimate are reported in Table 2, but all the abovementioned controls were included. Columns (2)-(5) in the table represent models with the same set of explanatory variables, with the difference being that the outcome variables are learner scores for each of the four literacy domains that formed part of the reading test.

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

The estimated treatment effect on the overall literacy score is 0.49 percentage points gained relative to the control group. However, we are unable to conclude with any level of statistical confidence that the true effect is statistically significantly different from zero. On the other hand, we are able to conclude 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 RCUP study initial aggregate analysis then showed that the intervention across randomly assigned schools had no substantial overall benefit for the learners in schools that received the programme.[xv] The intervention was not substantially beneficial, and therefore should not be rolled out across all demographically similar primary schools. If it was not effective in a relatively functional district with schools that were above average, we can reasonably infer that it would have same or even less effect for less functional districts and poorer or rural schools. The absence of an effect also requires further qualitative research in these primary schools, in order to identify reasons why the intervention did not effect any substantial change in performance. And so the research cycle continues.

Table 2 Main regression results

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

Note. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Discussion: External Validity Considerations in Designing Counterfactual Studies

The focus of this article has been on demonstrating the importance of appropriate sample sizes and a valid counterfactual estimate for producing a study with internal validity – i.e. the study has a logically sound basis for making a claim about the causal impact of a programme on the outcome of interest. Peters et al. (2015) suggest that internal validity is a necessary condition for a study to be regarded as having relevance for policy. They argue, however, that the sufficient condition for an empirical study is that it achieves external validity.

External validity refers to the transferability of a study’s results to the population as a whole or to a different population. This is often regarded as a weakness of RCTs due to the limited contexts in which they are often conducted and due to the artificially controlled study environment that is created (e.g. Pritchett & Sandefur, 2014). These concerns are certainly valid and are sometimes overlooked by those interpreting RCTs. In this section, we discuss several typical external validity challenges with education RCTs (especially in the South African context) and recommend steps that can be taken to (at least partially) mitigate the challenges.

Perhaps the most significant challenge to the external validity of an RCT relates to the context in which it was implemented. For practical and cost-related reasons, RCTs are usually implemented in a fairly small geographical area. For instance, the Reading Catch Up Programme discussed above was implemented in a single education district. The ‘Dr. Seuss question’ then emerges: if it had worked ‘here’, would it work over ‘there’, would it work ‘everywhere’? Underlying the question is the recognition that the impact of a programme may depend critically on specific contextual factors. For example, Pritchett and Sandefur (2014) have demonstrated that the impact of attending private schools will depend on the quality of public schools in the area. They similarly demonstrate that the impact of class size varies widely depending on the context, even amongst studies with a high degree of internal validity.

‘Context’ refers not only to geography, but to any relevant dimension, such as time. Lemons, Fuchs, Gilbert and Fuchs (2014) illustrate this through replicating five RCTs that had shown positive effects in the 1990s. The programmes (all targeting early reading instruction in Nashville, Tennessee) were no longer effective at improving learning outcomes relative to the control group when implemented in the 2000s. Lemons et al. (2014) maintain that the context had changed substantially – whereas previously reading instruction was not systematically incorporated into the curricula at kindergarten classes, these had since become institutionalised through various reforms. As a result, what had worked previously was no longer beneficial over and above existing instruction received at control schools.

A different set of external validity concerns relates to the special conditions created when implementing programmes in the context of an experimental research project. Perhaps the most well known of these is the ‘Hawthorne effect’. This occurs when the special attention given to the treatment group due to it being the subject of research contributes to changed behaviour and therefore improved outcomes. The impact evaluation is then confounded, because it is not possible to distinguish between the effect of the programme itself and the effect of the special attention. Related to Hawthorne effects are John Henry effects, in which the control group is affected through participating in the research. This means that the control group no longer represents a valid picture of the counterfactual. A small 'testing effect' is not fatal, as long as it affects the treatment and control groups equally. However, if the research component of the project substantially influences behaviour, it could mitigate the effectiveness finding of the programme itself (a false negative). As such, at-scale implemented policies would provide a better picture of the behaviour of schools, as neither treated nor control schools are under the impression that their behaviour will affect the outcome of the study. Unfortunately, once we look at purposive, at-scale implementation, we again face the evaluation problem and the need to find a suitable comparison group.

A somewhat different concern with RCTs is that the implementing partner is often an NGO or academic team, rather than those branches of Government that will ultimately implement a programme if taken to scale. These implementing partners may comprise small teams of highly competent and motivated individuals operating under a different accountability structure than will be the case when Government implements the programme. For example, Bold, Kimenyi, Mwabu, Ng'ang'a and Sandefur (2012) conducted an RCT of a contract teacher intervention in Kenya with two treatment arms and a control group. In one treatment arm the intervention was administered by an NGO, while in the other treatment arm the intervention was administered by the Kenyan government. Bold et al. (2012) found a large, statistically significant impact of the programme when implemented by the NGO, but zero impact when implemented by Government. This calls into question the external validity of many experiments that have shown a positive causal impact of a programme implemented by an effective NGO.

The usual concern around Hawthorne and scaling up effects is that the treatment effects will be inflated due to the special attention given to the treatment group. However, in the South African context, we have noted a potential bias in the opposite direction when evaluating systematic reforms. In order for an intervention targeting instructional change to be effective, there needs to be a high level of participation and enactment by teachers. For this to happen there needs to be strong support from the school principal, as well as district-level and provincial-level officials. There is a risk with an RCT (or any pilot initiative) that programme compliance will be low if teachers and principals do not perceive that the programme is a clear priority for the authorities. If this occurs, a programme may not show a positive effect in an RCT setting, but could indeed have an impact when Government implements it at scale. We call this problem the 'leadership factor': one does not want to provide too much special attention and leadership when conducting an RCT for fear of Hawthorne effects, but, conversely, programme effectiveness may require the system-wide leadership of the national, provincial and district authorities. Thus, these two forces would affect the RCT in opposite directions, with an uncertain net effect.

Another challenge to external validity occurs when a programme induces a different set of behaviours and effects when only some schools (or teachers or learners) are participating, compared to when all schools (or teachers or learners) are participating. For example, one way in which teachers can improve their instructional practices is through shared learning communities of teachers, often amongst small clusters of schools. In an RCT, however, these sorts of spill-over effects have to be prevented in order not to 'contaminate' the control group. It is thus possible that an instructional support programme could have larger benefits when teachers are able to share practices with each other, than when they are encouraged not to do so in an RCT. This we refer to as the 'spillover paradox'.

Although there are clearly a number of potential challenges to the external validity of RCTs, we do not believe that these concerns are serious enough to call the method into question. Rather, we recommend paying great attention to external validity considerations in the design and analysis phases of RCTs in order to mitigate the risks. Several steps can be taken to enhance the plausibility of making generalisations from the context of an RCT to a different or wider policy-relevant context.

First, choose a study population that is as representative as possible of the policy population of interest. For cost reasons, it may not always be possible to conduct an RCT in a nationally representative sample of students or schools. However, one can, for example, certainly avoid conducting an RCT in a particularly affluent or well-performing subset of schools.

Second, investigate heterogeneous effects, i.e. whether programme impact varies across relevant sub-groups. For example, in the evaluation of the Reading Catch Up Programme we found that programme impact was greater amongst learners with better initial levels of English proficiency. The strength of an RCT design is that subgroups are comparable across treatment and control groups. If we know for which sub-groups a programme was more effective, then we can come to make more educated conclusions about where else the programme could work.

Third, investigate intermediate outcomes along the causal chain. For instance, if a programme is intended to influence learner outcomes through changing teacher knowledge, then it is important to investigate whether teacher knowledge was indeed affected. It will be easier to transfer relevant lessons for policy and programme design when the generative mechanisms are understood, than when all we know is whether a programme was effective or not. Unpacking the causal black box is difficult, however, and requires an experimental design to allow researchers to perform a causal mediation analysis (see for instance Imai, Keele, Tingley & Yamamoto, 2011) and obtain further insights through qualitative research.

Fourth, design and evaluate interventions that do not rely on a model of implementation that is not feasible to operate on a large scale. The sample size requirements of an RCT – at least about 40 intervention schools are needed – implies a level of discipline on intervention design to prevent overly cost-intensive or expertise-intensive programmes being evaluated in the first place.

Fifth, work closely with Government at provincial and local levels to ensure high-level support of the programme that schools can perceive. But be careful to avoid spillover effects where Government officials begin to implement aspects of the programme in control schools.

Sixth, use innovative designs to avoid possible Hawthorne and John Henry effects. One may choose not to administer a pretest, for instance. With randomised assignment to treatment and control groups, there is no reason to expect any difference in outcome, except due to the causal impact of the programme. However, this will require a larger sample size, which may inflate costs substantially. Essentially, having a baseline measure and increasing the sample are two different ways of reducing sampling variance and improving power. One could also use routinely collected administrative data for outcome measures, where possible, to avoid special testing effects. Such data is unfortunately not always available or not of high enough quality. In South Africa, we do now have test score data for all schools in Grades One to Six and Nine (through the Annual National Assessments) and for National Senior Certificate candidates. For example, in our evaluation of the Reading Catch Up Programme, we complemented the analysis of independently collected test data with an analysis of the ANA data collected a few months after the programme ended.

Seventh, replication is a well-established mechanism for enhancing generalisability. Finally, there is a need to be realistic about the valuable yet limited and specific role of RCTs in the process of understanding school system improvement. Before an RCT is done, we do need smaller scale pilot studies to provide a proof of concept and to enhance programme design. Thereafter, an RCT could be warranted. If an RCT suggests a positive impact and is taken to scale by Government or an NGO with a wide reach, then we also need quasi-experimental research to evaluate the impact of the programme when implemented on a systemic scale. If large-scale roll-out is sequentially phased in, or targets a specific group of beneficiaries, there are possible quasi-experimental research designs with strong internal validity.

Conclusion

To advance education in South Africa we need to ensure that policy-makers and programme developers have access to genuinely trustworthy knowledge. There are signs that South African education leaders have become increasingly accustomed to (and have begun to rely on) evidence from cross-national survey studies like SACMEQ, TIMSS and PIRLS and evidence gathered from the Senior Certificate examinations and the ANAs. There is also growing awareness of the policy value of multivariate analyses, which consider correlations between variables of interest and educational outcomes (see for example Van der Berg, Girdwood, Shepherd, Van Wyk, Kruger, Viljoen, Ezeobi & Ntaka, 2014). However, the quest for estimates of the causal impacts of programmes and policies is most relevant to the policy-maker. Policy leaders therefore need to add to this list of findings from randomised control trials of education programmes and interventions, as well as causal estimates derived from quasi-experimental quantitative methods. The study of the Reading Catch-Up Programme study in Pinetown, Kwazulu-Natal shows the policy value of large-scale studies with rigorous estimates of the counterfactual, the strength of an RCT design. In this instance, having a valid estimate of the counterfactual prevented a false positive result. Millions of rands could have been spent on rolling out this programme across the province, only to discover later that it had little meaningful impact.

This is not to suggest that randomised control trials are the only or even the best approach to knowledge development in education in general or policy knowledge specifically. We recognise that in many instances, it is not feasible to subject a policy or programme to an impact evaluation, due to high cost and complex logistics. In such situations, well executed, qualitative studies will be indispensable, although it will still have to be recognised that a quantitative measure of impact will not be achievable through these methods. Although RCTs are costly, because interventions need to be carried out in at least 40 or 50 schools in order to satisfy statistical power requirements, there are often occasions where government does pilot an intervention or strategy on a large enough scale for an RCT, but neglects to roll out the intervention in a manner that facilitates the identification of a valid control group. This could easily be done through better planning, facilitated by collaboration between programme managers and evaluation specialists.

We recognise that the drive for impact evaluations can be construed as part of a centralist and technocratic tendency that pays less attention to locally developed innovations that may only be relevant in specific locales. Moreover, we recognise that good practice for evidence-informed policy would be a combination of rigorously designed randomised control trials complemented with equally rigorous qualitative case studies. The latter type of study can potentially provide real insights into the generative or change mechanisms (or the absence thereof). Complementary qualitative research can provide a fertile ground for piloting innovations and for developing research hypotheses that can be rigorously tested at scale using randomised control trials.

Notes

i. We would like to thank one of the anonymous reviewers for challenging comments on earlier drafts of the manuscript. We are suggesting that this replication study of a successful system-wide remediation programme vividly illustrates the value of a not widely used, but robust research methodology. The use of the randomised control trials, although not commonly used in South Africa, fits the criteria of a generally accepted research method. See for example two popular research texts - Cohen L, Manion L & Morrison K 2013. *Research methods in education* (7th ed). London, UK: Routledge and McMillan J & Schumacher S 2014. *Research in education: Evidence-based inquiry* (7th ed). Harlow, UK: Pearson Education Limited. In our view, the RCUP study illustrates the policy relevance of the RCT method for education systems characterised by transformation, and/or an emerging economy/development state, and/or scarce resources. RCT studies and related systematic reviews play a unique role in providing evidence to inform programme and policy implementation in the context of scarce state resources. On the critique pertaining to disconnect, the study's main finding i.e. gains in the intervention group, were equivalent to gains made by the control group, is possibly one of the most powerful illustrations of the value of large-scale impact evaluations with counterfactual components as it provided strong guidance on policy and/or programme adoption. While the literature review points to studies that show positive findings of effective literacy models, we observe that these studies make use of what we call 'illustrative' rather than 'substantive' counterfactuals. The RCUP study shows that when projects or programmes that show promise are subject to rigorous trials, trials with appropriate sample sizes and proper randomisation (schools and learners), we are likely to get a more accurate marginal impact estimate, that is, an estimate closer to the actual scale of impact likely with system-wide implementation.

ii. We think it is appropriate for these researchers to argue that the interventions were effective, but this must be done on the basis of personal observation, theory, deduction and argument, as opposed to being done on the basis of a statistical result. As case studies, these papers make a valid contribution. But confusion may result with the use of terms such as "control" group, and in the presenting of quantitative outcomes.

iii. Other quasi-experimental methods not discussed here include regression control, matching on observable characteristics, the use of panel data methods such as fixed effects, regression discontinuity methods, and instrumental variable methods. For a more detailed discussion see, for example, Duflo, Glennerster and Kremer (2006).

iv. However, one needs to bear in mind that very large samples can make any statistic estimated for the sample statistically significant. In such cases it is useful to investigate the confidence interval to establish if the effect size itself is meaningful.

v. The terminology of "treatment group" and "control group" originates from literature on medical trials, where a particular drug, or "treatment", undergoes trial. The terminology is now widely used across fields in impact evaluations. We use the terms "intervention group" and "treatment" group interchangeably.

vi. 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).

vii. Initially we tried to select schools based on the original below ANA 50% level, and between 30 and 90 learners criteria. But in order to find 100 schools we were obliged to start relaxing some of these criteria. Read the full sampling report in the pre-analysis plan to see the details of what we did.

viii. The power of the statistical test refers to the probability of avoiding a Type II error (that is, incorrectly rejecting a null hypothesis). Therefore it represents the likelihood of drawing the correct conclusions about the significance of differences between groups. Typically, a power level of 80% is considered high enough to detect differences, while keeping sample sizes reasonable.

ix. The ICC is the proportion of the total variation in test scores that is accounted for by between-school variation; the remainder is accounted for by within-school variation among learners. It describes the level of inequality between schools. The higher the ICC, the larger are the systematic differences in achievement scores between schools, and the more groups are required in the sample.

x. Having an equal sample size in the treatment and control groups is optimally efficient in achieving statistical power. However, as recommended by Duflo et al. (2006:30), when substantial costs are involved in implementing the treatment, a cost-effective solution can be to have a larger control group than treatment group. There is no statistical requirement for groups to be of equal size in order to be able to compare means or estimate coefficients in a regression model.

xi. In order to determine appropriate sample size, it is necessary to have some prior knowledge of the expected size of the intervention effect. In much of the contemporary US-based literature this has been standardised to a common effect size unit, that is, percentage of the standard deviation of the outcome measure. This allows for comparison across studies using different scales. While the original PRMP study did not report results in percentage of the standard deviation of the outcome measures, the percentage point gains reported were very high. The use of 0.2 standard deviations can be regarded as a moderate effect size relative to those typically observed in the international literature on school interventions.

xii. Given this core finding, the question of cost-effectiveness is of no consequence.

xiii. The stratification dummies refer to the characteristics according to which we stratified the initial sample of primary schools. These include an income quintile dummy (*high income quintil*e including Quintile 4 and randomly some Quintile 3 schools and low income quintile including Quintile 3 and Quintile 2 schools); smaller and larger schools dummy; and language performance in ANA 2013. For more detail, see detailed report on the sampling procedure which is available online in a pre-analysis plan on the RCT registry of the American Economic Association (https://www.socialscienceregistry.org/trials/405).

xiv. Although there is no reason to expect differences in endline test scores between the treatment schools and the control schools as an effect of causes other than the intervention, it is still worth including these other control variables, in order to enhance the statistical precision of the estimated treatment effect.

xv. When we disaggregated performance, some surprising insights emerged. These will be reported in a follow-up paper.

xvi. Published under a Creative Commons Attribution Licence.

References

Banerjee A, Cole S, Duflo E & Linden L 2007. Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, 122(3):1235–1264.

Bold T, Kimenyi M, Mwabu G, Ng’ang’a A & Sandefur J 2012. *Interventions & institutions: Experimental evidence on scaling up education reforms in Kenya*. Preliminary draft. Available at http://www.iies.su.se/polopoly_fs/1.101632.1348137980!/menu/standard/file/2012-08-06%20Kenya%20RCT.pdf. Accessed 6 January 2017.

Duflo E, Glennerster R & Kremer M 2006. *Using randomization in development economics research: A toolkit*. NBER Technical Working Paper 333. Cambridge, MA: National Bureau of Economic Research. Available at http://www.nber.org/papers/t0333.pdf. Accessed 29 December 2016.

Fleisch B & Schöer V 2014. Large-scale instructional reform in the Global South: insights from the mid-point evaluation of the Gauteng Primary Language and Mathematics Strategy. *South African Journal of Education*, 34(3): Art. # 933, 12 pages. doi: 10.15700/201409161040

Green JL, Camilli G & Elmore PB (eds.) 2012. *Handbook of complementary methods in education research* (4th ed). London, UK: Routledge.

Hellman L 2012. *GPLMS Intersen catch-up programme: Analysis of results*. Memo.

Imai K, Keele L, Tingley D & Yamamoto T 2011. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789. doi: 10.1017/S0003055411000414

Klapwijk N & Van der Walt C 2011. Measuring reading strategy knowledge transfer: Motivation for teachers to implement reading strategy instruction. *Per Linguam: A Journal for Language Learning*, 27(2):25–40. doi: 10.5785/27-2-106

Lemons CJ, Fuchs D, Gilbert JK & Fuchs LS 2014. Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, 43(5):242–252.

Mouton J 2013. *Report on evaluations of education improvement interventions*. Memo in files of Zenex Foundation.

Peters J, Langbein J & Roberts G 2015. *There’s no escape from external validity - reporting habits of randomized controlled trials*. Working Paper. Available at https://www.hhs.se/contentassets/95fb16b24d484cef8e8adda8e1f53f3e/peterslangbeinroberts_external-validity_20150313_draft.pdf. Accessed 5 January 2017.

Pretorius EJ & Lephalala M 2011. Reading comprehension in high-poverty schools: How should it be taught and how well does it work? *Per Linguam: A Journal for Language Learning*, 27(2):1–24. doi: 10.5785/27-2-105

Pritchett L & Sandefur J 2014. Context matters for size: why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2):161–197. doi: 10.1515/jgd-2014-0004

Van der Berg S, Girdwood E, Shepherd D, Van Wyk C, Kruger J, Viljoen J, Ezeobi O & Ntaka P 2014. *The impact of the introduction of grade R on learning outcomes*. Final report (Policy Summary, Executive Summary & Report Summary) for the Department of Basic Education and the Department of Performance Monitoring and Evaluation in the Presidency. Available at http://www.education.gov.za/Portals/0/Documents/Publications/Grade%20R%20Evaluation-1-3-25%20final.pdf?ver=2015-04-07-114113-503. Accessed 30 December 2016.