Failing to catch up in reading in the middle years: The findings of the impact evaluation of the Reading Catch-Up Programme in South Africa

Brahm Fleischa,*, Stephen Taylorb, Volker Schöerc, Thabo Mabogoaned

a Wits School of Education, University of the Witwatersrand, Johannesburg, South Africa  
b Department of Basic Education, Pretoria, South Africa  
c African Micro-Economic Research Unit, University of the Witwatersrand, Johannesburg, South Africa  
d Department of Planning, Monitoring and Evaluation, Pretoria, South Africa

A R T I C L E   I N F O

Article history:  
Received 16 June 2016  
Received in revised form 21 October 2016  
Accepted 29 November 2016  
Available online 7 January 2017

A B S T R A C T

In many developing countries, children are far behind the grade-specific curriculum. This article reports on an impact evaluation of one promising initiative, the Reading Catch-Up Programme, which was designed for a subset of South African schoolchildren, who are known to be behind the grade-specific English (second language) curriculum in the middle years. Overall, despite the promising findings of a preliminary non-experimental evaluation, the cluster randomised control trial study found no substantial or educationally meaningful programme impact. However, the study has identified two important secondary findings. First, for those schoolchildren with a stronger initial English proficiency, there was a moderate positive impact. This is despite the fact that the programme substituted the grade-level curriculum with remedial concepts that should have been covered in earlier grades. This suggests that the schoolchildren in the study may have been even further behind than was anticipated. A further finding from the impact evaluation relates to the effectiveness of instructional coaching as a component of the combined intervention model. The study found that the effectiveness of one-on-one instructional coaching may be dependent on the personal and professional characteristics of individual coaches. This finding complements existing research on the importance of the quality of the institutions implementing programmes.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Over the past decade there has been growing recognition that a substantial proportion of schoolchildren in low-income countries struggles to read and do basic mathematics. In India, for example, the 2014 ASER report suggests that half of the Standard V learners surveyed were reading at or below a Standard II level, and about one in five children were reading at a pre- or partial alphabetic level (ASER, 2014). Schoolchildren that are academically behind by the middle years of schooling are likely to fall further and further behind their counterparts as they progress up the school system. The high proportion of schoolchildren that are falling behind suggests that this is not just a problem of a small subset of children with specific organic or developmental barriers to learning, but is a challenge for the system as a whole. The widespread learning backlogs are systemic, affecting most students in the majority of schools in many poorly performing education systems.

How can education ministries and departments address these systemic learning backlogs? Although the achievement gap often starts in preschool and the early grades, the scale of the problem is only identified in the middle years, when the learner performance of education systems is benchmarked against that of other countries. And while there are a growing number of programmes designed to address early-grade reading and mathematics instruction, there are fewer programmes that have been developed specifically to address the systemic problem in the middle years of compulsory education. One exception is a system-wide initiative developed by the provincial government in the Gauteng province of South Africa. An eleven-week system-wide remedial programme targeted at Grade 4–6 students in over 1000 schools (about 70% of all public schools) aimed to reteach reading and writing in English as a second language. An interval evaluation of this system-wide initiative (Hellman, 2012) suggested that it was highly effective in helping the majority of underachieving Grade 4–6 students improve their basic English-literacy proficiency.
Although these results were clearly encouraging, the design of the evaluation was not rigorous. The internal evaluation was undertaken by the service provider that had designed the intervention, the pretests and the posttests were administered by the teachers involved in the initiative, and the study contained no counterfactual.

Given the need for evidence of cost-effective system-wide catch-up or remedial programmes, particularly in low-income countries with uneven levels of academic performance, the research team decided to retest the impact of the Reading Catch-Up Programme (RCUP), using a cluster randomised control trial design. Although positive evidence generated from small-scale studies (such as Pretorius, 2014) has contributed to the knowledge base on effective interventions, the design features of many of these types of studies have limited the generalisability of the findings of the studies. Cluster randomised control trials, by design, have the advantage of strong internal validity, and they are able to provide robust estimates of effect sizes.

Against this backdrop, this article describes the background, context, design, and findings of the Reading Catch-Up Programme. We do this in six sections. Following this brief introduction, which situates the study in an international context, the second section provides a detailed description of the research method, with an explanation of the intervention, the cluster randomised control trial (RCT) methodology, the rationale for the selection of the study site, and the data-collection processes. This is followed by a presentation of the major findings of the study, which draw on baseline and endline testing, as well as an analysis of the 2014 Annual National Assessments for the treatment and the control school groups. We also present disaggregated findings, which point to the differential effect of the intervention, and what we have come to call a “coach effect”. The article concludes by contextualising these findings within the growing body of knowledge about early-grade reading and system-wide change.

The primary aim of the study was to address the question about the relative effectiveness of the Reading Catch-Up Programme and the model that it represents, but it also included sub-questions that allowed the researchers to understand more nuanced aspects of the programme.

The main research question in the pre-analysis plan was:

1. Does the RCUP have a positive average impact on Grade 4 learners’ endline evaluation scores (literacy test)? Was this positive average impact consistent across all subtasks of the literacy test?

The sub-questions included:

1. Are students, teachers and schools with varying characteristics likely to experience differences in the magnitude of impact of the RCUP on learning outcomes?

2. Will the effect of treatment differ depending on the individual doing the coaching?²

2. Literature review

What does the literature tell us about effective remediation programmes that close the literacy achievement gap in the middle years of schooling across education systems? While there is limited literature that specifically uses the term “catch-up programme” or “catch-up intervention”, there is a small but growing body of literature that addresses interventions for “students who struggle academically” (this phrase is taken from Richards-Tutor et al., 2015). And while the empirical literature remains largely based on studies undertaken in the Global North, particularly in North America and the United Kingdom, there are a handful of studies in lower-middle-income and low-income countries that provide an important base on which to build policy and programme insights.

Torgesen (1998) identified four critical elements to be considered in evaluating the effect of system-wide remediation programmes, namely (a) the appropriateness and the quality of the instruction provided, (b) the intensity and the duration of the programme, (c) the fit in terms of the needs of the students, and (d) the timing of the intervention. With regard to the first element, he argues that all programmes, irrespective of whether the programme is intended for students that are achieving in line with curriculum expectations or whether it is for students that are struggling academically, should meet the criterion of being “structured, systematic and explicit”. In addition, he argues that all programmes, particularly “preventative” or “remediation” programmes, should incorporate phonemic awareness, letter-sound correspondence, blending skills, pronunciation conventions, strategies for multisyllabic words, and automatic recognition of high-frequency “irregular” words, all within a structure that encourages meaning-making in reading and writing. On the intensity and the duration of the programmes, Torgesen argues that these programmes would generally require substantially higher levels of intensity than conventional classroom instruction, as expressed in smaller classes or one-on-one instruction. On the third element, he argues that it is essential to ensure that the programmes match the particular needs of the students served. Without an adequate understanding of the specific weaknesses (and strengths) that the students experience, it is likely that inappropriate programmes will be selected. Finally, Torgesen noted the importance of identifying the most effective grades or phases in which to begin interventions.

To understand the findings of the empirical literature, it is useful to begin with a summary of the findings of recent systematic reviews. In a review of studies on the effectiveness of remedial programmes for English-language learners, Cheung and Slavin (2012) found that programmes that showed consistently positive results had in common the use of extensive professional development, coaching, and cooperative learning. More recently, Richards-Tutor et al. (2015) undertook another research synthesis of intervention programmes for English learners of English (i.e., second-language speakers) that are effective in improving reading performance. They found 12 peer-reviewed studies on remedial programmes for “at-risk” learners, or learners of English. Seven of these studies found moderate to large effect sizes for early-grade interventions. Five studies showed similar effect size gains in comprehension only. In addition to effect sizes, the Richards-Tutor et al. (2015) review also explored issues related to intervention intensity, group sizes, duration, personnel, and quality. Siddiqui et al. (2015) recently evaluated the effectiveness of two particular programmes designed to remediate literacy levels of underachieving students. In an evaluation of the “Switch-on Reading” programme, Gorard et al. (2014) found that a ten-week literacy intervention for Level 4 students had an effect size of +0.24. Siddiqui et al.’s (2015) evaluation of the twenty-week “Accelerated...
Reader” programme also had an effect size of +0.24. The Switch-on Reading programme made use of a widely used Reading Recovery approach, while the Accelerated Reader programme used a web-based intervention that monitors and manages students’ reading practice and encourages independent reading. The Reading Recovery programme has consistently shown positive results, from D’Agostino and Murphy’s (2004) meta-analysis to the most recent RCT study, namely May et al. (2015).

What does the research say about tightly designed and scripted literacy programmes? A recent systematic review undertaken by Snilstveit et al. (2015) found that interventions developed to improve content, subjects or topics, that are tightly aligned around learning materials and teacher training, had the largest and consistent positive effects on learning outcomes. While there has been a strident critique of scripted lesson plans and principled objections to them within the teacher education community, this element of the instructional infrastructure is receiving growing recognition. Beaty (2011) reminded the research community that many of the key progressive education reform movements of the past, particularly Montessori’s early childhood instructional programme relied heavily on tightly scripted lesson plans to fracture old patterns of instruction and put in place newer approaches and methods. Reeves (2011), while only partially supportive of scripted (2013) plans, shows how teachers moving into a new domain of teaching benefit from designed scripted lesson plans to learn the new teaching practice. Most recently, Hiebert and Morris (2012) has argued persuasively that detailed lesson plans, what they refer to as ‘annotated’ lesson plans are a key resource or set of artefacts to guide the new practice that move the debate from a focus on teachers to a focus on teaching.

The growth in the number of various different effective interventions, in both the early grades and at the middle-school level, is certainly encouraging, but these interventions are unlikely to speak to very different settings in the Global South. Rather than a relatively small proportion of English-language learners or at-risk students mainstreaming into an overwhelmingly well-resourced English-learning environment, the challenge for interventions in the Global South is to demonstrate that these interventions can work at scale for the majority of students in challenging classroom settings in the context of fragile education systems. Possibly the most promising of such effective interventions have been undertaken in India. Glewwe and Muralidharan (2015) have summarised four key studies of what they refer to as supplemental remedial schooling interventions. The first study they describe, namely Banerjee et al. (2005), is an impact evaluation of a pull-out programme that specifically targeted the weakest readers in Grades 3 and 4 in schools in Mumbai and Vadodara. This initiative showed positive gains one and two years after the intervention. In the Banerjee et al. (2010) study of the impact of community participation, they found that a youth volunteer programme of 2–3 months in which low-paid volunteers provided remedial teaching in holiday camps improved students’ performance in reading and mathematics, and that the gains persisted after two years. The authors suggest that the main reason for the success of the programme was that volunteers taught at the level of the students, rather than attempting to comply with the unrealistic expectations of the official curriculum. Lakshminarayana et al.’s (2013) study also evaluated the impact of using community volunteers who provided remedial lessons. Unlike the earlier programme study, however, the volunteers provided two hours of tutoring in the afternoon in whatever subject extra support was needed. The Lakshminarayana et al. (2013) study found a large impact in this after-school remedial instructional programme. The most recent study, namely Banerjee et al. (2015), is the Pratham multi-treatment randomised control trial of these remedial instruction approaches within public schools. As in the earlier studies, Banerjee et al. (2015) found that teaching children at the level that matches their learning, by either volunteers or teachers after school, was effective. However, when ordinary public school teachers used the method within the normal school day, it had no impact. The reason for this finding, according to both Glewwe and Muralidharan (2015) and Banerjee et al., (2014), is that ordinary teachers in the normal school day are bound to complete the syllabus and keep pace with the state textbook, regardless of their students’ actual literacy or mathematics skill levels.

Within the South African context, a recent small-scale catch-up programme study found that strong positive gains can be achieved with an extended resource-intensive intervention. Pretorius’s (2014) study describes an intervention programme that evolved to focus on the basics, which included intensive work over a year with a literacy coach for four days per week in a single school, and support in building up substantive resources. The study showed that there were substantial gains made by the students, but, in contrast to the study conducted in India, the gains were highest for the strongest students, as was evident from the scores on the baseline test.

While both the Indian and the South African studies are clearly suggestive of what is possible, the knowledge base of rigorously designed interventions that allow system-wide catch-up remains very limited. Given the scale of this specific challenge, there is clearly a strong need for more and better research, to ensure better evidence for informed policy decision-making.

3. Background to the reading Catch-Up programme

In 2011 the Gauteng Department of Education, under its Gauteng Primary Language and Mathematics Strategy, developed and implemented a reading remediation programme in 792 underperforming primary schools. What came to be called the “Catch-Up Programme” contains three key elements, popularly called the “education triple cocktail”, which included a set of scripted lesson plans, provision of high-quality reading materials, and on-site one-to-one instructional coaching.

The schools that were part of the initiative were instructed to suspend normal teaching of the curriculum for the term, and to replace it with a remedial curriculum covering English First Additional Language topics that should have been covered in Grades 1 to 3. This remedial curriculum was recontextualised (in the Bernsteinian sense) through daily scripted lesson plans covering 11 weeks of teaching.

The intervention provided six sets of teaching and learning resources. The teachers received printed black-and-white lesson plan guides and assessment record books. The classes were given four listening and teaching posters covering the themes “In the Classroom”, “At the Zoo”, “On the Beach” and “At the Hospital”, and sets of English “readers” intended for the early grades. Each student received two special exercise books, one to write in during the regular class time during the intervention, and a second specifically for tests.

The designers of the Catch-Up Programme prescribed a strict and consistent weekly teaching regimen, to be followed according to the same routine every week. The teaching week was divided into seven half-hour teaching periods.

The scripted lesson plans and teacher and student resources were seen as necessary but not sufficient to remediate the students’ poor reading skills. The intervention also provided just-in-time training and ongoing in-class coaching to each teacher. The instructional coaches played a number of roles in the programme. They provided training to the teachers in small groups, and they visited the classes to model teaching practice. They observed, supported, and provided encouragement as the teachers enacted
the lesson plans. The coaches also monitored and tracked teachers’ compliance.

What was the change theory that underpinned the intervention? The theory of change in which the programme was embedded was built on a number of key assumptions. The first assumption was that the lesson plans and the coaching would change how instructional time would be used and understood. The first page of the lesson plan guide clearly linked particular lessons to specific calendar days, thus specifying the pace at which the learning programme unfolded. The pace remained the same even if teachers were absent or the day was interrupted for any reason. The responsibility, or burden, shifts to the teacher to keep up with the pre-specified timeframes. Within the lesson itself, teachers needed to increase their stamina to keep up with the forward motion of the activities. The role of the coaches was to assist the teachers, and once trust had been established, to push them harder, so that they remained on track and keep up. What the new use of time was intended to do was to increase the amount of time spent on learning tasks, and to intensify work on the tasks, thus allowing for increased opportunities to learn, and effective coverage of the curriculum. The prescribed weekly lesson routine, or regimen, created a defined structure to school and lesson time.

The second assumption was that the lesson plans and the learning resources were intended to expand the teachers’ pedagogic and classroom management repertoire. The posters and related lesson plans offered the teachers with new insights into techniques to increase vocabulary and use of the early-grade readers to improve reading fluency. The change theory assumed that while teachers may have had some exposure to these techniques at one time or another, the programme provided a systematic, structured, and sequential framework for their enactment. The lesson plans also afforded tangible instruction on the organisation of time and resources, i.e. classroom management.

The third assumption was that the lesson plans would re-orient instruction to the actual levels of the students, rather than the levels assumed by the official curriculum. An emerging finding in international literature on large-scale reform is the negative consequences of an overambitious curriculum (Pritchett and Beatty, 2012). By beginning at the students’ actual reading levels and skills, and moving the students systematically along, the intervention would ensure that the majority of the students would be able to benefit from the instruction and the reading materials.

The intervention was implemented in the second term of 2012 in 792 public primary schools in the Gauteng province. The finding of the internal evaluation was positive. The report focused on student performance, assessing the extent to which the programme had improved spelling, language, comprehension, and writing skills between the pretest and the posttest. Two assessment instruments were developed, one for students in Grades 4 and 5, and another one for learners in Grades 6 and 7. The final test sample consisted of 1570 classes, and included 45% of the teachers that participated in the programme.

The evaluators assigned the learners into four quartiles on the basis of their pretest scores (see Fig. 1), and then calculated the average gains for each quartile on posttest. The programme seemed to have had a strong, positive, and consistent effect (Hellman, 2012).

While the average scores offered important findings regarding the magnitude of improvement for the sample, an analysis of the distribution patterns yields additional insights into the effectiveness of the programme.

The literature review revealed that improvement of this magnitude and scale is unusual. If the results could be corroborated, and the intervention successfully replicated in a different context, this would represent an important contribution to the knowledge base on educational change.

4. Research design

To test the efficacy and transferability of the Gauteng remediation programme, and to do so in a way that would ensure robust and rigorous findings, a cluster randomised control trial design was used.

The Pinetown district of the KwaZulu-Natal Department of Education was selected as the research site. It had the advantage of containing a range of poor schools of different types (rural, urban, informal, and formal). At the time of the study, the funder was engaged in a larger intervention in the province, which was aimed at improving language and mathematics performance in primary schools in the district. As such, the Reading Catch-Up Programme would have provided useful insights.

Particular care was taken in designing the most appropriate sampling frame and sample size for the study, to ensure optimal statistical power, as well as to satisfy ethical and cost concerns. 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).

Initially, we tried to select schools based on the original sub-50% ANA level, and the 30–90 learners criterion. But in order to find 100 schools, we had to start relaxing some of these criteria. Read the full sampling report in the pre-analysis plan, to see exactly what we did.
In addition to measuring the immediate impact of the intervention on average grade reading performance, we used official data from national standardised tests (referred to as Annual National Assessments) as a second, albeit slightly delayed, measure of achievement.

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

Fig. 2 shows the distribution of baseline scores (expressed as percentage scores) for learners in both the treatment schools and the control schools. The figure indicates how similar the distributions of performance were between the treatment and the control schools, which confirms that the randomisation procedure that was followed was successful in producing adequate balance between the two groups. Fig. 2 also shows that the vast majority of the learners scored below 20% on the pretest. Given the very low scores on the pretest, concerns were raised about a possible “floor effect”. This may have had the unfortunate effect of making it harder to identify improvements in learning at the bottom end of the distribution.

Overall, of the 2663 learners that wrote the pretest, 2466 completed the posttest, which represents a 7.4% attrition rate. The attrition rate was slightly higher in the control group than in the treatment group. When running a regression to test whether allocation to the treatment group predicts attrition, it is evident that treatment does not predict attrition at all once variables such as baseline scores are controlled for. Therefore, we excluded from the data set those learners that were absent, and we proceeded to analyse the data using only learners that were present in both the pretest and the posttest.5

5 For information on pretest results and details of implementation, please consult the Zenex Report.

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

57% 51%
56% 46%
23% 0%
20% 6%

L1 L2 L3 L4

Fig. 1. Catch-Up Programme distribution across the four quartiles, 2012.

Source: Hellman (2012)

Fig. 2. Kernel density of pretest scores, percentage.

The trend lines in the pretests and posttests for the treatment and the control schools shows that while both groups improved substantially between the pretest and the posttest, the improvement is only marginally better in the treatment group. In other words, while the baseline trends were very similar, so were the endline trends.

The small difference in improvement in the treatment schools relative to the control schools is clear upon observation of Fig. 4. Although the posttest score was slightly higher in the treatment schools than in the control group, the difference is not statistically significant.

Table 1 shows the results of five regression models. 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 baseline or pretest score of the student, stratification dummies, the gender of the student, their age, exposure to English at home, frequency of an adult reading at home, class size, the age of the teacher, their gender, their qualifications, and school size. Although there is no reason to expect differences in endline test score between the treatment schools and the control schools as an effect of causes other than the
treatment effect. Only the coefficient on the treatment variable and the standard error of the estimate are reported in Table 1, but all the above-mentioned controls were included. Columns (2)-(5) in the table represent models with the same set of explanatory variables, with the difference being that the outcome variables are student scores for each of the four literacy domains which 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 an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools. “Effect in SD” refers to the estimated treatment effect as a proportion of the standard deviation amongst endline test scores within the control group.

The estimated treatment effect on the overall literacy score is 0.49% points gained relative to the control group. This amounts to an estimated effect size of 0.024 standard deviations, which is not of an educationally meaningful size, and we are unable to conclude with any level of statistical confidence that the true effect is statistically significant different from 0. On the other hand, we are able to conclude with high levels of statistical confidence that the intervention improved spelling outcomes and language (including grammar) outcomes for learners in the treatment schools. We estimate that spelling improved by 1.27% points relative to the control group, and that language improved by 3.96% points. The estimated impact on comprehension and writing items was not statistically significantly different from 0.

6. Impact on annual national assessments

The Annual National Assessments were written during the week of 16–19 September 2014 across schools in South Africa. The Annual National Assessment (ANA) are government standardised tests administered to all students in Grades 1 to 6 and 9 in South African public schools to measure Home Language, English First Additional Language, or Afrikaans First Additional Language and Mathematics. The universal ANA tests are administered and marked by classroom teachers in the October of each year. In Pinetown, 94% of learners in our sample of treatment and control schools wrote the English as First Additional Language ANA test.7

The use of ANA data for this study has several advantages. Firstly, it provides an alternative data source, and one which is less likely to have been affected by any Hawthorne effect. Secondly, it reflects educational performance a little while after the intervention was completed, thus indicating whether any impacts persisted. Thirdly, it allows the use of all learners in treatment and control schools, that is, not only those learners who were sampled for our own independent testing. This provides us with a data set of 6419 learners across our treatment and control schools. While this improves the statistical power to identify a treatment effect, the disadvantage of this approach is that we do not have a baseline score for each learner. The best we can do is to control for each school’s average ANA score in previous years.

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

---

7 There are, however, several limitations of the ANA data for our purposes. The quality of the data is not expected to be as high as that of the data collected by our service provider. This is because the ANA tests were locally administered and marked by teachers within each school. Differences in the conditions of testing, and in marking standards across schools, should make the data a somewhat noisy signal of learner proficiency. This is confirmed by the respective correlations between our baseline test score, our endline test score, and the ANA language scores of learners. In a sample of 1928 learners who we were able to match between the RCUP and ANA data sets, the correlation coefficient between the baseline test score and the endline test score was 0.86. However, the correlation coefficient between baseline score and ANA English score was only 0.53, and between endline score and ANA English score it was 0.56. Noisy data would be expected to cause a degree of attenuation bias in the estimated treatment effects (where the estimated effect is biased towards 0). Fortunately, though, there is no reason to expect differences in marking, or quality of ANA information, to be correlated with assignment to treatment.
of a large learning disadvantage for males in schools in South Africa. Fig. 5 shows the distributions of test scores for learners in intervention schools and learners in control schools. The distributions indicate that learners in intervention schools scored higher than learners in control schools. This is a preliminary indication of a positive treatment effect.

The first hypothesis to test was whether students in intervention schools performed better in the Grade 4 English ANA test than did students in control schools. Table 2 reports the results of several regression models estimating the impact of the RCUP on ANA data. When no attempt is made to control for baseline differences in achievement, the estimated treatment effect is 3.35% points, and this is statistically significant at the 90% level.\(^8\) Models 2, 3, 4 and 5 in Table 3 show the estimated treatment effect when different ways of controlling for prior school performance are used (namely controlling for school mean language score in ANA 2013, controlling for school mean language score in ANA 2012, and controlling for school mean language score in both ANA 2012 and ANA 2013). In all cases, the estimated treatment effect is somewhere between 3 and 4% points, but in models 4 and 5 it is not statistically significant.

Notes: All models include controls for stratification dummies, learner gender, and learner age. Standard errors are adjusted for the fact that learners are clustered in schools. Since a few schools wrote English as Home Language, only 91 schools are represented in the models (37 treatment schools, and 54 control schools). The results are robust to an alternative specification where the outcome variable is percentage score, irrespective of whether the score was from the English as Home Language test or the English as First Additional Language test. “Effect in SD” refers to the estimated treatment effect as a proportion of the standard deviation amongst endline test scores within the control group.

7. Coach effect

Did the impact of the intervention depend on which coach the school was allocated? The service provider used two coaches to implement the programme. Each coach was allocated 20 schools. Therefore, one can estimate two separate treatment effects, one for each coach. Table 3 shows the results when running the exact same regression models as reported above, but instead of including a single-treatment dummy variable, we include two dummy variables (one for each coach), still relative to the reference category of control schools. There are two main limitations in this analysis. The first limitation is that the coaches were not randomly assigned to schools. However, the fact that we have baseline scores for each student, and therefore can control for stratification and other student, school and teacher characteristics, reduces the likelihood of omitted-variable bias. The second limitation is that the effective sample size is cut in half: instead of having a treatment group of 40 schools, we now compare each treatment group of 20 schools to each other, and to the control group. This means that standard errors will be larger, making it less likely to observe a statistically significant treatment effect.

Table 4 shows no significant impact for “Coach B” on any of the outcomes. For “Coach A”, however, there were statistically significant effects on both spelling and language. The coefficients for Coach A are all larger than in the overall treatment effects, as reported in Table 4 (though we cannot conclude with statistical certainty that the effects are larger). Table 4 confirms the same result, this time using the ANA data. In both Tables 4 and 5 the size of the estimated effect for Coach A is substantial – up to 0.28 and 0.23 standard deviations, respectively. Therefore, this provides suggestive evidence that the success of an intervention that uses coaches to support teachers may depend on the particular person doing the coaching. If indeed this was the case, we are not able to determine what characteristics of Coach A led to a larger impact.

Notes: The reference category for both coaches is the control group. All models include controls for baseline score, stratification dummies, student gender, student age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications, and school size. Standard errors are adjusted for the fact that learners are clustered in schools. “Effect in SD” refers to the estimated treatment effect as a proportion of the standard deviation amongst endline test scores within the control group.

Notes: The model includes controls for school mean language score in ANA 2012 and ANA 2013, stratification dummies, learner gender, and learner age. Standard errors are adjusted for the fact that learners are clustered in schools. “Effect in SD” refers to the estimated treatment effect as a proportion of the standard deviation amongst endline test scores within the control group.

8. A “Matthew effect”

The phenomenon where initially stronger children often benefit more from additional support programmes or educational resources has sometimes been referred to as the “Matthew effect”. We find evidence of this with respect to the impact of the RCUP, and we illustrate this using several analytical and descriptive methods. Our analysis of whether baseline achievement mediated the treatment effect is actually one of several possible “heterogeneous effects” that we tested for – that is to say, we tested whether the impact of the programme was different depending on various student, school or teacher characteristics. There was no evidence of heterogeneous effects based on student gender, student age, student exposure to English at home, or class size (full results are not reported here).

Fig. 6 provides convincing evidence of a Matthew effect. Amongst the bottom four deciles of baseline achievement, it would seem that learners in control schools actually gained more than learners in treated schools. In the top five deciles, however, learners in treated schools improved their performance more than learners in control schools.

Fig. 7 provides a similar picture, only now the level of smoothing is greater, thanks to the use of Lowess smoothing lines. The same pattern is evident. Amongst the weakest 40% of
The impact of the coaches (ANA language).  

| Coach A | 5.38** (2.45) |  
| Effect in SD | 0.229 |  
| Coach B | 0.85 (1.70) |  
| Effect in SD | 0.036 |  
| N | 6419 |  
| r² | 0.2106 |  

Table 4: The impact of the coaches (ANA language).  

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

Impact by baseline performance of learners.  

| Combined score | Spelling | Language | Comprehension | Writing |  
| Treatment | −0.44 (0.86) | 0.32 (0.7) | 2.92** (1.15) | −1.96 (1.54) | 1.39 (1.47) | 0.49 (0.04) | 0.03 (0.04) | 0.28 (0.06) |  
| Baseline percentage score | 0.97 (0.02) | 0.93 (0.02) | 0.66 (0.03) | 0.79 (0.03) | 0.53 (0.04) |  
| Treatment x Baseline | 0.05 (0.04) | 0.05* (0.03) | 0.07 (0.04) | 0.02 (0.04) |  
| N | 2466 | 2466 | 2466 | 2466 | 2466 |  
| r² | 0.77 | 0.77 | 0.46 | 0.53 | 0.28 |  

Note: *p < 0.1, **p < 0.05, ***p < 0.01.
subtests where positive treatment effects were observed. The various estimated treatment coefficients are graphically represented in Fig. 8. For spelling and language there is evidence of larger effects in the upper parts of the performance distribution.

9. Discussion

The findings above clearly show that students’ English literacy skills gains cannot really be attributed to participation of schools in the Reading Catch-Up Programme. Even though the increases in students’ spelling and language scores in the treatment schools are statistically significant, and the ANA scores show statistically significant relative gains compared to the control groups, the gains may have limited educational meaning. The achievement level difference between the treatment schools and the control schools, as measured by standardised scores, is relatively modest, particularly when compared to the strong positive findings reported in the 2012 reading catch-up study, and in Pretorius’ (2014) study. A review of students’ posttest scripts from treatment schools clearly shows that most of the Grade 4 students continue to be very weak spellers, with limited command of the basic structures of the language, and with little reading comprehension and writing proficiency. The gap between these students’ literacy performance and the demands of the curriculum remains large.

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

10. Explaining the limited gains in the treatment group

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

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

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

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

While the main finding shows little real difference in gains between the treatment and the control groups of schools overall, the more nuanced analyses provide important insights into the possible conditions under which meaningful change, what Hopkins (2001) has described as “improvement for real”, could occur. An added insight that emerges, one that will require new studies to confirm, is the differential impact of individual coaches. The RCUPE findings suggest that while instructional infrastructure (Cohen, 2011), in the form of lesson plans, learner resources, and coaches, may be necessary conditions for improvement, the quality and effectiveness of individual coaches may be an often hidden but powerful factor.

Besides this strong finding, the study has also provided substantial evidence relating to a range of themes. These include further evidence of the serious underperformance in English as a First Additional Language at the start of the middle years, and the scale of the performance gap between the genders. The study pretest data set suggested that the Grade 4 learners’ English-language proficiency is very weak. Pinetown was selected as one of the higher-performing districts in the province, as indicated by the ANA scores. Our findings, however, suggest that there is a significant discrepancy between performance levels, as indicated by the ANA scores, and proficiency levels, as measured by our test. The divergent performance measures may be a function of the different test instruments, or of the different conditions under which the tests were administered and marked.

Another major insight from the pretest analysis is the large performance gap between boys and girls. This gap is evident in both the pretest and the posttest, and is consistent across the study tests and the ANA results. This trend, identified by Perry and Fleisch (2006) in the late 2000s, and recently confirmed by Zuze and Reddy (2014), is not adequately understood.

11. Conclusion

As the evidence mounts on what works to improve early-grade reading (e.g. Snilstveit et al., 2015; McEwan, 2015), policymakers are increasingly aware of the need of another school group, namely the millions of children in the higher grades that have missed out on receiving improved teaching of reading and mathematics in the early grades. Policymakers, particularly in the Global South, are increasingly asking questions of the research community about what works to close the gap between the expectations of the national curriculum and the low levels of reading and mathematics skills of students that have been in school for three or more years and who have missed out. While there is growing consensus that programmes such as Reading Recovery are effective (D’Agostino and Murphy, 2004; May et al., 2015), particularly if implemented in the first years of schooling, other than the well-known Indian studies of supplemental instructional practices outside classrooms (Glewwe and Muralidharan, 2015), little is known about cost-effective interventions that can catch up cohorts of students across entire systems.

This study provides evidence from a randomised control trial (RCT), to start to fill this gap in the literature. Specifically, the trial was designed to provide estimates of the impact of a catch-up programme that was shown to be successful in a similar setting in South Africa. Unlike the original pretest-posttest study report, we found little evidence from the Pinetown RCT study that the Reading Catch-Up Programme had an educationally meaningful impact. The Catch-Up Programme study did, however, find evidence of a differential impact on students, advantaging students with higher initial baseline scores. It also showed some evidence of a differential coach impact. These findings are consistent with and can extend the existing literature, and they will be pursued in more focused articles in the future.

What might account for the absence of positive impact for the Reading Catch-Up Programme? In terms of Torgeson (1998) implementation framework i.e. (a) the appropriateness and the quality of the programme, (b) the intensity and the duration of the programme, (c) the fit in terms of the needs of the students, and (d) the timing of the intervention, a number of observations can be made. Although we cannot comment on the quality of the programme, there is so evidence that the programme may have been too short given that it took some time for teachers to become accustomed to the new programme. While the schools that participated in the intervention all implemented the programme with some degree of fidelity, it is plausible that an intervention of this kind requires a long lead time for some teachers to really begin to own and enact the practice as intended. The question of duration is important as it is possible that the actual amount of time, namely 10 weeks, was simply too short for students to benefit from the new practice. Our secondary analysis of the students’ misspellings of monosyllabic three-letter words on the spelling test component suggests that most of the students had not mastered the rudimentary elements of English phonics, particularly the 22 English vowel sounds. It is possible that the timing of the programme, implementation in the second quarter of the school year might have impacted on its effectiveness, although it is unlikely as students and teachers have settled by this time of the year and learning fatigue has not yet begun. In terms of Torgeson framework, with regards to the fit with the needs of the students, the intervention might not have been appropriately “pitched” at the learning levels of the students. While the intervention did “re-teach” the early-grade literacy skills, its starting point might have been too high, and the pace and progression was too rapid. This would be consistent with the explanation offered by Pritchett and Beatty (2015).

While the 2012 Catch-Up Programme study reported strong student gains and suggested that many of the middle-year students in Gauteng “caught up” with key reading skills, the absence of similar educationally meaningful impact results in the 2014 Reading Catch-Up Programme RCT precludes a policy warrant for this programme. With these findings, policymakers are cautioned against adopting both the programme and the model as they currently exist. Should the programme or model be modified, by extending the duration of the programme, slowing its pace and progression, improving the quality of the coaches, and reducing the coach-teacher ratios, it may be that a modified programme may show the promise of the 2012 implementation. But from a policy perspective, it would be prudent to retest the model, again using robust counterfactual methods.

While this study has clearly demonstrated that the RCT is a useful tool for showing “what works” or in this case “what does not work”, the approach is limited in that it provides only incomplete or partial insights into the contextual, institutional, personal and professional elements that makes change possible. This is made clear from the findings of the coach effect. To better understand the generative mechanism of change, qualitative case studies are required to uncover deeper and possibly unknown factors and the individual characteristics of what makes coaching effective. This certainly provides a strong motivation for complementary research, combining RCTs with rigorous qualitative case-study research.

Acknowledgements

The research from which this article draws was supported by the Zenex Foundation. We would like to thank the top management of the KwaZulu-Natal Department of Education, for agreeing to host
the study, and the district officials and teachers in the Pinetown
district that consented to participate in the study. This article was
language-edited by a freelance language editor, Anthony Sparg.

References

Publications/ASER%20Reports/ASER%202014/National%20PO%20PPTs/
from Two Randomized Experiments in India (No 11904). National Bureau
of Economic Research, Cambridge, MA.
of participatory programs: evidence from a randomized evaluation in education in
Banerje, A.V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Walton, M.,
2014. Teaching at the Right Level: Evidence from Randomized Evaluations in
India. Draft Paper.
Beatty, B., 2011. The dilemma of scripted instruction: comparing teacher autonomy,
fidelity, and resistance in the freobelian kindergarten, montessori, direct
Cheung, A.C., Slavin, R.E., 2012. Effective reading programs for Spanish-dominant
English language learners (ELLs) in the elementary grades: a synthesis of
0034654312465472.
Cambridge, MA.
Glewwe, P., Muralidharan, K., 2015. Improving School Education Outcomes in
Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. RISE
Working Paper 15/001.
Gorard, S.A.C., Siddiqui, N., See, B.H., 2014. An evaluation of the ‘Switch-on Reading’
Hiebert, J., Morris, A., 2012. Teaching, rather than teachers, as a path toward
Lalchiminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D., Mann, V.,
2013. The Support to Rural India’s Public Education System (STRIPES) trial: a
cluster randomised controlled trial of supplementary teaching, learning
material and material support. PLoS One 8 (7), e65775.
May, H., Gray, A., Sirindes, P., Goldworthy, H., Armijo, M., Sam, C., Gillespie, J.N.,
Tognatta, N., 2015. Year one results from the multisite randomized evaluation of
McEwan, P.J., 2015. Improving learning in primary schools of developing countries:
Perry, H., Fleisch, B., 2006. Gender and educational achievement in South Africa. In:
Town, pp. 107–126.
Prentorius, E.J., 2014. Supporting transition or playing catch-up in Grade 4?
Implications for standards in education and training. Perspect. Educ. 32 (1), 51–
76.
258.
effectiveness of reading interventions for English learners: a research synthesis.
Siddiqui, N., Gorard, S., See, B.H., 2015. Accelerated Reader as a literacy catch-up
intervention during primary to secondary school transition phase. Educ. Rev. 1–
16.
Snijstvest, B., Stevenson, J., Phillips, D., Vojkova, M., Gallagher, E., Schmidt, T., Jobse,
H., Geelen, M., Pastorolio, M.G., Eyers, J., 2015. Interventions for Improving
Learning Outcomes and Access to Education in Low- and Middle-income
Countries: A Systematic Review. 3ie Final Review. International Initiative for
Impact Evaluation (3ie), London.
Torgesen, J.K., 1998. Catch them before they fall: identification and assessment to
Zuze, T.L., Reddy, V., 2014. School resources and the gender reading literacy gap in