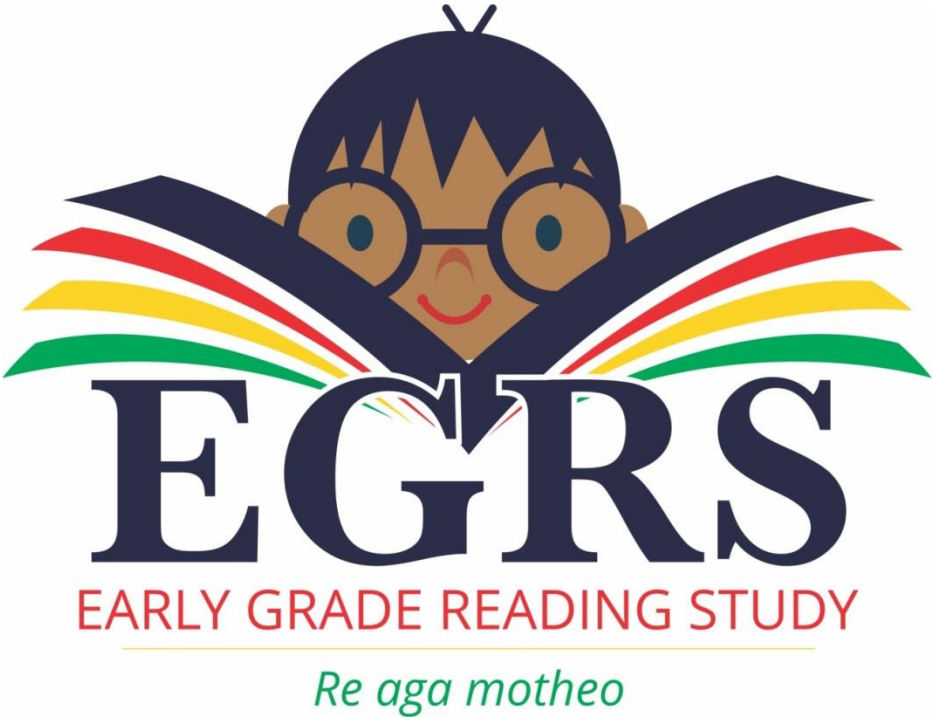
THE EARLY GRADE READING STUDY:

Impact evaluation after two years of interventions

Technical Report



IMPROVING EARLY GRADE READING IN SOUTH AFRICA [P2.10.SA.IE]

13 OCTOBER 2017

Authors: Stephen Taylor* (Department of Basic Education), Jacobus Cilliers (Georgetown University), Cas Prinsloo (Human Sciences Research Council), Brahm Fleisch (University of the Witwatersrand) & Vijay Reddy (Human Sciences Research Council)

*Corresponding author: Stephen Taylor – tayor.s@dbe.gov.za

The Early Grade Reading Study team gratefully acknowledges the generous support from the donors and partners listed below.

Education and Sport Development
Department of Education and Sport Development
Departement van Onderwys en Sportontwikkeling
Lefapha la Thuto le Tlhabololo ya Metshameko
NORTH WEST PROVINCE

basic education
Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

planning, monitoring & evaluation
Department:
Planning, Monitoring and Evaluation
REPUBLIC OF SOUTH AFRICA

ZENEX FOUNDATION

unicef
unite for children

2030 NDP

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

EGRS
EARLY GRADE READING STUDY
Re aga motheo

AngloAmerican

class act

Read to Lead
A Reading Nation is a Leading Nation

3ie International Initiative for Impact Evaluation

HSRC
Human Sciences Research Council

CONTENTS

ABBREVIATIONS AND ACRONYMS

DBE	Department of Basic Education
PIRLS	Progress in International Reading Literacy Study
RCT	Randomized Controlled Trial
ANA	Annual National Assessments
EGRS	Early Grade Reading Study
SGB	School Governing Body
RDD	Regression Discontinuity Design
EGRA	Early Grade Reading Assessment
HSRC	Human Sciences Research Council
CRC	Community Reading Coach
GPLMS	Gauteng Primary Language and Mathematics Strategy

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The Department of Basic Education would like to acknowledge the work of the EGRS Research Team who provided oversight to the project and to the production of this report. The Research Team includes Stephen Taylor (DBE), Brahm Fleisch (University of the Witwatersrand), Jacobus Cilliers (Georgetown University), Cas Prinsloo (HSRC) and Vijay Reddy (HSRC). Carol Nuga-Deliwe, Nompumelelo Mohohlwane and Janeli Kotzé have worked closely with the Research Team and have played a critical role at an intellectual and direction-setting level.

This project would not have been possible without generous financial support from a number of partners. In this regard we would like to thank the International Initiative for Impact Evaluation (3ie) for funding the evaluation side of the project and for providing peer reviewed feedback and support throughout the project. From 3ie, there was significant advice and support provided by Howard White, Beryl Leach, Deo-Gracias Houndolo and Heather Lanthorne. We would also like to thank the ZENEX Foundation, UNICEF and Anglo American Chairman's Fund for funding the implementation of the three EGRS interventions. In particular, we thank Gail Campbell, Fatima Adam and the rest of the ZENEX Foundation staff team for engaging on the content of the project and providing critical feedback and support. Similarly, at UNICEF we would like to thank Saadhna Panday-Soobrayan, Nadi Albino and Wycliffe Otieno for robust discussions and support on the evaluation findings at various stages throughout the project.

The DBE would like to thank Class Act Educational Services for doing an excellent job of delivering the three EGRS interventions. The quality of training provided and responsiveness to additional requests went above and beyond what was formally required. In particular, we would like to acknowledge the work of Penny Groome, Lorraine Marneweck, Maryna du Plooy, Nonhlanhla Hlomuka, Thili Thela and Kaitlin Alsofrom at Class Act, as well as the Coaches and Trainers – Dorothy Molapo, Sabi Mlambo, Helen Kgobane and Kgomotso Phaletse.

The Human Sciences Research Council (HSRC) have done a tremendous job of managing the evaluation side of the project - data collection, data capturing and working with the research team on instrument design and research interpretation. Led by Cas Prinsloo, we must thank the HSRC team members: Lolita Winnaar, Charlotte Motha, Matthews Makgamatha, Jacqueline Harvey, Shawn Rogers, Elmi de Koning and Matselane Maja. For Setswana language expertise on the development and revision of learner tests, we acknowledge the excellent work done by Badiri Moila. Social Data Collectors, led by Rosina Langa, were subcontracted by the HSRC to conduct fieldwork in Waves 2 and 3 and were a pleasure to work with.

We would also like to thank the North West Department of Education for partnering with us as the host province. The hard work by Mpho Motlhabane, Mahlodi Teu, Thabo Phorabatho, Godi Poti and Boitumelo Moeng is much appreciated. We also recognize the support from officials in the districts of Ngaka Modiri Molema (led by the Director Benny Monale) and Dr Kenneth Kaunda (led by the Director Hasimbhay Motara) together with all the curriculum coordinators and area office staff who supported the project and served on the curriculum reference group.

The Department of Planning, Monitoring and Evaluation has been a hugely supportive partner, both financially and through including EGRS on the National Evaluation Plan. The personal involvement of Thabo Mabogoane, Ian Goldman, Jabu Mathe and Christel Jacob has been much appreciated.

We also thank the University of the Witwatersrand for undertaking the financial management and contracting regarding the interventions side of the EGRS project as well as for academic collaboration, led by Brahm Fleisch.

Members of the DBE Directorate for Research Coordination, Monitoring and Evaluation have served tirelessly to coordinate and manage this project. We thank Justice Libago, Sasabona Maseko, Tshirado Tshirado, Neo Makgato, Ernest Rasekgwalo, Tariro Chiweza, Mikateko Mbambo and Jesal Kika for their hard work.

Within the DBE, certain individuals across various parts of the organization have supported the project and provided advice along the way and deserve special mention: Mathanzima Mweli, Waheeda Carvello, Haroon Mahomed, Chintha Maharaj, and Martin Gustafsson.

A number of other people have advised and supported the EGRS from inception and continue to do so. We thank Servaas van der Berg, Luis Crouch, Kamilla Gumede, Linda Zuze, Cheryl Reeves, Ursula Hoadley, Nicholas Spaull, Nick Taylor, Kerryn Dixon, Bobby Soobrayan, Benjamin Piper, Sonja van der Vyfer and Saguna Gordhan.

We are grateful to the teachers, the parents and ultimately the children who participated in this project in various ways. We hope this project has benefited them but also recognize that their participation is intended to lead to the benefit of children throughout the system going forward.

EXECUTIVE SUMMARY

PROJECT OVERVIEW

One of the biggest developmental challenges facing South Africa is the high number of children who do no learn to read for meaning in the early years of school. This is the foundational skill upon which all others build and as such this has become a leading priority for the Department of Basic Education (DBE). In order to address this challenge, the DBE initiated the Early Grade Reading Study (EGRS) in collaboration with academics at the University of the Witwatersrand, the Human Sciences Research Council (HSRC) and Georgetown University (USA). This is a large-scale educational impact evaluation – the biggest in South Africa - and aims to build evidence about what works to improve the teaching and learning of early grade reading in African languages in the country.

The core of the project is a comparison of the cost-effectiveness of three promising intervention models to improve reading outcomes in learners' home language (Setswana). The project commenced in 2015 by working in 230 quintile 1-3 schools in the North West province. Each intervention has been implemented in a separate group of 50 schools with a further 80 control schools where ordinary schooling is continuing. The project uses a formal impact evaluation methodology known as a Randomised Control Trial (RCT) complemented with a 60-classroom observation study and eight detailed case-studies. The study design enables the researchers to estimate the impact of each intervention model on measures of reading, as well as understand where, how and why different elements of the intervention models are working.

The evaluation assessed three interventions:

1. **A structured learning programme & centralised training**: The first intervention provides teachers with lesson plans aligned to the National Curriculum Statement Grades R-12 (NCS) including the Curriculum and Assessment Policy Statements (CAPS), as well as additional quality reading materials and training at centralized workshops twice a year.
2. **A structured learning programme & specialist on-site coaching**: The second intervention (implemented in a different group of 50 schools) provides teachers with the same set of lesson plans and reading materials but provides ongoing support to teachers through on-site coaching and small cluster training sessions.
3. **Parental intervention:** The third intervention (implemented in a further 50 schools) holds weekly meetings with parents to discuss the importance of learning to read in the early grades and to empower them with the knowledge and tools to become more involved in their child's literacy development.

The three interventions were implemented in the grade 1 class of 2015 and at the grade 2 level in 2016, thus following the same cohort of learners. This year (2017) the two structured pedagogic interventions have continued at the grade 3 level, thus ensuring that this cohort of learners were exposed to the interventions for the entire Foundation Phase.

Three waves of data have been collected to date. A baseline data collection ("Wave 1") was collected at the start of 2015 when learners had just begun grade 1. A midline data collection ("Wave 2") was collected at the end of 2015. A third wave of data was collected at the end of 2016, when most learners were in grade 2. Data collected towards the end of 2016 when the learners had received two years of the interventions forms the basis for the evaluation findings presented in this report.

YEAR 2 EVALUATION FINDINGS: WHAT WORKS?

Of the three intervention models we have been evaluating, the Coaching intervention is showing a substantial positive impact after two years of intervention (end of grade 2). This intervention included lesson plans, reading materials and on-site coaching by reading experts. **Learners who received two years of this Coaching intervention were approximately 40% of a year of learning ahead of the students in the schools that received no intervention ('business-as-usual' schools).** This is a truly significant improvement by international standards. The other two interventions (centralised Training; and the parent involvement intervention) appeared to have a small positive impact, less than half the size of the coaching intervention.

We measure the impact of the Home Language Literacy interventions on children letter recognition, word recognition, non-word recognition, paragraph reading (oral reading fluency), phonological awareness, comprehension, writing and two additional school subjects, English and mathematics – in case there were spillover effects.

Although the Training intervention had moderate positive effects on some of the sub-tests, the Coaching intervention registered statistically significant positive effects on all home language literacy measures, with similar effect sizes across the sub-tests. There was no significant effect of the coaching intervention on the short mathematics test that was administered. This means that we have no evidence of a negative effect through crowding out of teaching time for mathematics. Interestingly, we observe a significant positive effect on English. This might be attributable to an improved underlying language ability (obtained through the home language intervention) or simply due to improved classroom management and transferable instructional methods acquired by the teacher through the coaching intervention. Either way, this is an encouraging finding for the Coaching intervention.

Although the overall impact of the parent intervention was small, it does appear to have had a significant positive impact on phonological awareness. This was probably the specific reading skill that was most directly targeted through the parent meetings. Sound games were a key

method taught to parents to use at homes in the development of their child's phonological awareness.

WHO BENEFITS MOST FROM THE INTERVENTIONS?

Boys catch up somewhat: The effective Coaching intervention is helping boys catch up some of the way to girls. Although girls still perform better than boys in the "Coaching" group, the gap is smaller than it is in the control group.

Impact concentrated in urban schools: For all three interventions, the observed impacts are larger in urban township settings, but there is no measurable impact in deep rural settings. This means that we may need to approach interventions in rural schools differently.

Middle-to-top learners benefited most: The impact of Coaching is largest for children in the middle and upper part of the achievement distribution with small or negligible impacts for the weakest children. Importantly, there is no evidence of a negative effect for any part of the performance distribution. One implication of this finding is that structured pedagogic programmes that make use of lesson plans may benefit certain groups of children more, depending on the level at which the lessons are set.

Large-classes benefited most: Both the teacher support interventions ("Training" and "Coaching") had the largest impacts in relatively large classes (38 to 45 learners). In smaller classes, it may be that teachers in the control schools are already able to effectively manage classrooms, provide structured learning and differentiated attention to a variety of learners. However, in larger classes the EGRS interventions helped teachers to provide better instruction in a challenging environment. Both of the pedagogic interventions emphasized good classroom management practices such as how to reorganize classrooms, work in small groups while keeping the larger classroom occupied and bring routines and predictability to the classroom. However, in the very largest classes (50 plus learners) the impact of EGRS interventions was smaller, possibly indicating that beyond a certain threshold it remains difficult to conduct effective teaching. This emphasizes the need to eliminate excessive class sizes (50+) in the Foundation Phase.

HOW MUCH DID TEACHING PRACTICE AND PARENT BEHAVIOUR SHIFT IN RESPONSE TO EGRS INTERVENTIONS?

Through the use of mixed methods research (teacher questionnaires in all 230 schools, lesson observations in 60 schools and a set of detailed case studies), we investigate underlying change mechanisms by observing how the learning environment, teaching practice, and classroom activities changed as a result of the programmes.

If any teacher support programme is to be effective, teachers themselves need to feel positively inclined towards the programme. Based on questionnaires administered to teachers in all

groups of schools, teachers in the Coaching schools were considerably more likely to report feeling a high level of professional support than those in the control schools, with teachers in the Training group also somewhat more likely to experience high professional support.

Two other results are worth emphasizing. First, even though there is no large difference in access to graded readers, the lesson observations reveal that **far more learners are actually reading graded readers in the Coaching and Training schools**. **This increase is substantially larger for teachers who received Coaching relative to teachers who received Training**. Second, even though we find no change in the probability *that* learners practice reading in the classroom, there is a noticeable difference in *how* they practice reading: **Teachers in both Training and Coaching groups are more likely to do group-guided reading, resulting in more opportunities for learners to receive individual attention. The impact is, again, larger for teachers who received Coaching relative to Training**. These results suggest that there are some teaching practices such as group-guided reading that are difficult to enact and require additional development to be effective. They also reveal an important interaction between resources and teaching practice: graded readers are only useful if teachers have developed the skills to use them effectively in the classroom.

Low attendance was a major limitation in the Parent intervention. In 2015, just over a third of parents attended at least three sessions while in 2016 just under a third attended at least three sessions. Nevertheless, compared to the control group, parents in this intervention group reported attending a significantly higher number of parent meetings at their school. However, no other indicators of parental involvement in home reading or educational activities shifted substantially, confirming that there was no large change in parental behavior in response to the intervention.

RECOMMENDATIONS AND DBE PLANS FOR MOVING FORWARD

1. **Structured programmes with coaches help**: A structured learning programme aligned to the NCS, together with additional high quality reading support materials (graded reading books, flash cards, posters), can make a significant difference to learning outcomes, if accompanied by effective and carefully monitored support to teachers (coaches).
2. **Coaching is the best alternative**: Whereas previously very little evidence existed about effective large-scale teacher support modalities in South Africa, we now have evidence that on-site coaching to Foundation Phase teachers can shift learning outcomes, and that this is a cost-effective strategy. Modelling of lessons, in a safe space, as they navigate the lesson plans for teaching learners to read is critical.
3. **Direct in-service training better than 'train-the-trainer' models**: Direct in-service training of teachers (4 two-day workshops over the course of 2 years), while less effective than on-site coaching, is in turn likely to have more impact than "cascade" models where specialists "train the trainers" who then interact with teachers.
4. **Existing subject advisers cannot fulfil the role of a coach**: The low ratio of subject advisors to schools (especially in the Foundation Phase) makes it impossible for subject

advisors to fulfil the role of reading coaches, as implemented in EGRS; nor do we recommend increasing the number of subject advisors to allow this since the recruitment process, oversight structures and modus operandi of the coaches is different to that of subject advisors.

5. **Prioritizing schools is a viable option**: On-site coaching interventions could be implemented in priority schools (e.g. 100 or 500 schools in a province) on a temporary basis (e.g. 2 years at a time) and through independent contracting and oversight structures. The cost for 100 schools would be R6 million at current prices.
6. **Develop reading norms in the African languages**: Reading norms cannot simply be adapted from one language into another due to differences in language structures. It is a complex exercise requiring longitudinal data. Therefore, the EGRS data could be used towards the development of reading norms in the African languages.
7. **Learning from EGRS**: Other large scale intervention initiatives such as those administered by the NECT could draw on the lessons of the EGRS and extend successful programmes to selected schools and districts.
8. **Parental involvement needs further research & may be promising:** Whilst parental involvement is a hugely deterministic factor in a child's learning outcomes, the biggest challenge from a policy perspective is how to shift parent involvement at scale. Given the potential cost-effectiveness of such interventions, researchers and policy-makers should continue to investigate mechanisms to do so.
9. **Learning what works in deep rural settings:** Formative research and subsequent impact evaluation is required to figure out what kinds of school support programmes make a meaningful difference in deep rural settings.
10. **Measuring long-run EGRS impacts:** The DBE is planning to administer subsequent data collections on the same sample of learners to measure the long run impacts of these reading interventions.
11. **EGRS for EFAL in Mpumalanga**: A second phase of the Early Grade Reading Study (EGRS 2) is underway in the Mpumalanga province, since the start of 2017. This project aims to investigate the effectiveness of two alternative interventions on English as First Additional Language in the Foundation Phase.
12. **Early Grade Mathematics Study**: We also hope to conduct an Early Grade Mathematics Study over the next 2 to 5 years, with the first activity being a detailed scoping study to identify and design promising interventions with strong theories of change as well as cost-structures that would be sustainable on a large scale.

1. CONTEXT: WHAT MOTIVATED THE EARLY GRADE READING STUDY?

The acquisition of reading is foundational to all subsequent learning; yet the majority of South African children are being left behind in this regard. South Africa's participation in international assessments of reading and literacy has revealed that the majority of children in grades 4, 5 and 6 have not yet learned to read with comprehension.[1] These children, who have not learned to read, can thus not read to learn in subsequent grades and in all their subjects.

Reading outcomes vary dramatically by socio-economic status and by province. For example, the pre-PIRLS study of 2011 showed that 83% of grade 4 children in Limpopo could not read for meaning compared with 27% in the Western Cape. Consequently, massive inequalities in educational achievement are established early in primary school. Spaull *et al* (2016) estimate that by grade 2, more than half of children in Quintile 1 to 4 schools[2] are already not "on track" relative to curriculum expectations.[3] Sadly, there is no evidence of these inequalities being reduced in later years. Therefore, early interventions, such as improving the acquisition of reading amongst poor children, can be expected to have larger effects than interventions later in the school programme.

> Perhaps the most effective way to bring about socio-economic transformation in South Africa is to improve reading outcomes amongst poor children.

Based on the available evidence and basic reason, one can therefore conclude that weak reading foundations are the root cause of the dropping out of school that occurs amongst 16 to 18 year-olds in South Africa. Since performance in the secondary school leaving examination, known as "matric", is strongly predictive of post-schooling success and labour market success, and wage inequality is the main driver of overall economic inequality in South Africa, it is no exaggeration to suggest that the most effective way to bring about meaningful socio-economic transformation in South Africa is to improve the learning and teaching of reading in schools serving less affluent and historically disadvantaged communities.

[1] The Progress in International Reading Literacy Study (PIRLS) of 2006 indicated that more than 80% of grade 5 children in South Africa had not yet learned to read with meaning and, according to Kathleen Trong (2009: 104), were at "serious risk of not becoming literate". The pre-PIRLS study of 2011 confirmed that the majority of our grade 4 children were substantially behind the grade-appropriate reading level. The SACMEQ study of 2007 indicated that a large proportion (even by the standards of the region) of grade 6 South African children were still "functionally illiterate".

[2] Public schools in South Africa are classified into five "poverty quintiles" based on the socio-economic characteristics of the surrounding community, though these are not in fact equal numbers of schools in each quintile. School funding norms take quintile status into consideration.

[3] Taylor (2013) analysed data from the National School Effectiveness Study and found that children in historically black schools were the learning equivalent of three grade levels behind those in historically white schools by the time they were reached grade 5.

Although there have been and are various initiatives underway to support early grade reading, there is little or no sense of what is working and why. Intervention programmes lead by the DBE, provincial departments or other partners such as the National Education Collaboration Trust (NECT), are typically not set up with an impact evaluation design in mind. Moreover, there are competing models of support in the system. Some initiatives use so-called cascade training models where district officials are orientated to a particular programme or new set of resources, and they in turn train teachers on a decentralized basis. Sometimes, conventional teacher training workshops are held. Since 2010, the Gauteng Primary Literacy and Mathematics Strategy (GPLMS) provided additional graded reading booklets and scripted lesson plans to teachers. The programme also featured regular on-site visits from specialist reading coaches to observe classroom practice and offer assistance.

There is, however, a lack of evaluation presenting plausible causal estimates of the impact of these initiatives. One innovative paper used a quasi-experimental method to estimate the impact of the GPLMS programme, and found a substantial positive effect (Fleisch and Schoer, 2014). However, there were methodological limitations to this paper given available data. It is important that a national reading strategy be based on scientific evidence regarding what most improves the acquisition of reading. The EGRS was designed in part to provide evidence about the potential effectiveness of some of the models and change mechanisms that are being employed in similar initiatives around the country.

A randomized controlled trial (RCT) design allows a credible estimation of the causal impact of interventions, and thus has the potential to inform responsible policy decisions. By using a lottery to allocate schools to intervention and control groups it is possible to construct a credible "counterfactual" scenario – what would have happened to those who received an intervention had they not received that intervention.

Truly random assignment to intervention and control groups means that any differences in outcomes observed after the intervention period can be *attributed* to the impact of those interventions.

Moreover, by directly comparing the impacts on reading outcomes of alternative programmes, each with different cost implications, we can identify the most cost-effective intervention. This project is designed to explicitly compare the impact and cost of a new model of teacher development (on-school support) to the impact and cost of a more traditional model (training at central venues). The third intervention, which aims at improving parent involvement in schools and in home-based reading activities, relies on a rather different theory of change and is less expensive. By measuring the success of each intervention on the same scale, this project will provide a sense of the cost-effectiveness of different policy options.

The primary implementing partner is the South African government, in particular the Department of Basic Education. A key role is also being played by the North West provincial education department, which is contributing financially and is championing the project within the schools.

A service provider has been appointed to run the three interventions on behalf of the DBE for the purposes of this impact evaluation. The service provider is an organisation called "Class Act", which is highly involved in partnerships with government to run literacy interventions. For example, "Class Act" was a service provider in the Gauteng Province's implementation of the GPLMS programme over the last few years. Programme interventions are being funded by a coalition of donors, including the ZENEX Foundation, UNICEF, Anglo American and the Department of Planning, Monitoring and Evaluation in the Presidency. These funds are being managed by the University of the Witwatersrand, which ran a tender for the service provider work and subsequently entered into a contract with Class Act.

The evaluation side of the project is being supervised by the Research Team while the data collection and capturing is being managed by South Africa's Human Sciences Research Council (HSRC). The evaluation is being funded by the International Initiative for Impact Evaluation (3ie).

2. LITERATURE REVIEW: EARLY GRADE READING INTERVENTIONS LOCALLY AND INTERNATIONALLY

Improving the quality of education in developing countries has been a conundrum that has received significant attention over the past two decades. Having conducted extensive research describing weak learning outcomes, researchers in South Africa and around the world are shifting their focus to identifying solutions – policies and programmes that will lead to improved results. Consequently, there has been a surge in the implementation of various interventions aimed at affecting change, and commensurately so, the evaluation of these interventions. In this pursuit of improving the quality of education, a broad range of interventions have been tried and evaluated, including the provision of information about the quality of schooling to parents, providing in-service training to teachers, providing additional resources to schools, providing new technology to classrooms, teachers or learners, implementing nutritional and health interventions, changing the incentive and accountability structures in which teachers operate, and cash transfers or merit-based scholarships. However, the question remains as to which of these groups of interventions are most effective in affecting change.

The first generation of school based experiments mainly focused on increasing school resources, but found that these inputs did not produce the expected achievement gains. The provision of improved school infrastructure has been found to affect school attendance, but has not made a significant impact on learner performance (Adukia, 2017; Freeman, et al., 2011). Learning and teaching support material (LTSM) are often assumed to be essential in supporting learning, however, the mere distribution of these resources has not proven to be successful in improving learning outcomes (Glewwe, Kremer, & Moulin, 2009; Glewwe, Kremer, Moulin, &

Zitzewitz, 2004; Das, et al., 2011; Sabarwal, Evans, & Marshak, 2014). The effectiveness of resources such as LTSM appears to depend on how well these are used by teachers, and even by factors such as whether learners are able to read those materials. These studies, however, have mainly focused on the distribution of resources, without additional support or training in the use of these resources.

A different strand of interventions focused on improving learner health and nutrition with the purpose of enabling learners to attend school more regularly and to learn more productively. Nutrition and health interventions are thought to affect learning outcomes through improving learner attendance and learner concentration. These interventions typically include programmes that administer either deworming medication or nutritional supplements to learners, running a school feeding scheme that provides a nutritious meal to learners, the provision of reading glasses, providing immunisations and malaria prevention programmes. School feeding programmes have shown to have a small effect on learning outcomes, but a larger effect on learner attendance especially in areas where food security is low (Diagne, Lô, Sokhna, & Diallo, 2014; Ismail, Jarvis, & Borja-Vega, 2012; Altman, 2013; McEwan, 2013). The evidence of deworming programmes is largely inconclusive, with some evaluations of deworming finding some educational benefits for learners (Ebenezer, et al., 2013; Miguel & Kremer, 2004), but evaluations in other contexts not finding any positive impact (Watkins, Cruz, & Pollitt, 1996; Simeon, Grantham-McGregor, Callender, & Wong, 1995). Similarly, programmes that aim to provide micronutrient supplements to learners had beneficial effects on learners in some contexts (Luo, et al., 2012; Wong, Shi, Luo, Zhang, & Rozelle, 2014), but not in others (Jukes, Zuilkowski, Parawan, & Lee, 2014).

The lack of accountability of schools and teachers for professional conduct and providing quality education is often cited as a reason for weak learner performance in developing countries. High levels of teacher absenteeism and the fact that teacher pay is typically unrelated to any performance measures, for instance, are often cited as evidence of this. This has led to experiments with interventions that aim to improve education outcomes through parents or communities holding schools accountable. Similar to the health and nutrition interventions, the results of increasing accountability measures vary greatly and seem to be largely dependent on the context. Positive results have been found for interventions that provide parents and community members with more information on the oversight role that they can play, or that provide parents with school score cards (Andrabi, Das, & Khwaja, 2015; Pandey & Goyal, 2011). However, in different contexts the provision of information to parents did not lead to any increased community involvement or teacher effort (Banerjee, Banerji, Duflo, Glennerster, & Khemani, 2010; Nguyen & Lassibille, 2008).

A significant body of research has shown that teachers and their teaching are critical to learner performance (e.g. Hanushek, 2010). Various interventions therefore aim to improve the quality of teaching, either through teacher training, teacher incentives, changes to how teachers are hired or providing diagnostic feedback to teachers. Evidence seems to suggest that teachers do not necessarily change their instructional practices in response to financial incentives, but that

change may occur if participation in a training programme has explicit implications for promotions or salary increases (Glewwe, Illias, & Kremer, Teacher incentives, 2010; Muralidharan & Sundararaman, 2011; Popova, Evans, & Arancibia, 2016). In a systematic review of 26 teacher in-service training programmes, Popova et al. (2016) found that the largest impacts on student learning come from programmes that focus primarily on classroom management.[4] However, their results also suggest that more generic training programmes without any focus on a specific subject have a lower impact on student learning (Popova, Evans, & Arancibia, 2016). This review also showed that face-to-face training sessions at universities or training centres had larger impacts on learner performance than sessions held at centralised venues such as government buildings or hotels, although the former programmes may typically also be more time-intensive. Programme impact was also found to be improved if follow-up visits that review the material that was taught in the initial training was included in the programme. Researchers or government officials were also found to be less effective trainers than education practitioners (Popova, Evans, & Arancibia, 2016).

Programmes which seem to have consistently shown positive and large impacts on learner performance are programmes where resources are provided as part of a more comprehensive intervention package which includes training and support in the use of these resources (Nonoyama-Tarumi & Bredenberg, 2009; Piper, Zuilkowski, & Mugenda, 2014). Similarly, the review by Popova, et al. (2016) also showed that the largest impacts on student learning were found in programmes that provide certain materials alongside the teacher training. In a systematic review conducted by Snilstveit, et al. (2016), programmes that combine teacher training and resources are classified as structured pedagogy programmes. They define the central element of structured pedagogy as “the development of evidence-based curricula and instructional approaches, along with lesson plans and training for teachers in delivering new content and materials for students” (Snilstveit, et al., 2016: 25). In their review, Snilstveit, et al. identified 21 different structured pedagogy programmes and found that these programmes are effective in improving learner perfomance in both mathematics and reading, but that larger impacts were observed in the language programmes. Studies that found particularly high gains include the Primary Maths and Reading Rural Expansion programme in Kenya and the School Readiness Programme in Cambodia (Nonoyama-Tarumi & Bredenberg, 2009; Piper, Zuilkowski, & Mugenda, 2014).

In South Africa three studies measuring the impact of reading programmes on learner performance are worth noting. The first programme, in 2000, was the Learning for Living project which was implemented by the Read Educational Trust (READ). The programme provided resources to classrooms and mentoring to teachers. The evaluation took on a quasi-experimental design, where READ staff members nominated schools that had achieved high levels of programme implementation as the treatment schools to be evaluated. Government officials further selected schools that closely mirrored the demographics of the intervention schools as the control schools. This design was of course unlikely to sufficiently control for

[4] The authors of the review, however, state that this finding is driven by a single program, the In-Service Teacher Education programme in Thailand (Nitsaisook & Anderson, 1989)

selection effects and remains a limitation. The evaluation was conducted four years into the implementation and assessed Grade 1 and Grade 2 learners in their home language and found strong positive effects for the READ Home Language Initiative (Sailors, Hoffman, Pearson, Beretvas, & Matthee, 2010).

The second study piloted a programme called the Systematic Method for Reading Success (SMRS), which used both lesson plans and the Early Grade Reading Assessment (EGRA) tool, was conducted by the Molteno Institute of Language and Literacy in collaboration with the Department of Education, and was evaluated by RTI International. The programme was conducted in ten treatment and five control schools in each of the Limpopo, Mpumalanga and North West provinces in South Africa during 2009. The evaluation found that over a five month period the programme increased the average letters read per minute from 1.75 per minute to 16.09 per minute (Piper, 2009). The small sample size of this study, however, makes it difficult to gain a good understanding regarding the scalability of this programme.

The final study on early grade reading in South Africa is the Gauteng Primary Literacy and Maths Strategy (GPLMS) that was implemented in the Gauteng province in 2010. The core components of the strategy included daily lesson plans, high-quality learning and teaching materials and ongoing instructional coaching. The study was evaluated using a Regression Discontinuity Design, but suffers from various limitations in the identification of a control group and with the outcomes data. Notwithstanding these limitations, both the implementation and the evaluation of GPLMS suggested that a high quality structured learning programme supported by instructional coaching could be effective at a relatively large scale (Fleisch B. , Schöer, Roberts, & Thornton, 2016).

Given the ever increasing body of research that has been conducted in search of programmes which affect change in learner performance, numerous systemic reviews or meta-analyses have been conducted in recent years focusing on identifying effective types of education interventions (Conn, 2017; Glewwe, Hanushek, Humpage, & Ravina, 2014; Krishnaratne, White, & Carpenter, 2013; McEwan, 2012; Murnane & Ganimian, 2014; Snilstveit, et al., 2016; Kremer, Brannen, & Glennerster, 2013). It has, however, emerged that although the various systematic reviews all claim to summarize the findings of robust studies on this question, the conclusions they come to are different. Evans and Popova (2016) discuss the discrepancies between the various conclusions of the multiple systematic reviews as well as the methodological differences that led to the varying conclusions. Regardless of these methodological differences, Evans and Popova (2016) found that across all the reviews it is evident that (i) teacher training interventions will be most effective when tailored to the teacher involved; (ii) pedagogical interventions must change students' learning experiences and be adapted to individual student learning levels; (iii) teacher training may be most effective when it is repeated and linked to a specific pedagogical method or tool; and (iv) increasing accountability can also improve student learning.

Evans and Popova (2016) themselves haven't been without their critics. Haddaway, et al. (2015) have criticized them as having used methods that are at wide variance from accepted

criteria for systematic reviews or synthesis. These authors join other strident critics of both Randomised Control Trials (RCTs) and systematic reviews. Deaton and Cartwright (2016) have a number of criticisms of RCTs, but possibly the most useful is their argument that these studies seldom explore the mechanisms that actually make the interventions work. Without understanding these mechanisms, RCTs have limited policy utility as it difficult to make an assessment of what can and cannot be transferred or scaled up. Nadel and Pritchett's (2016) critique is more or less in line with Haddaway, et al. (2015). They argue that including different interventions that appear to be in the same 'class' or 'type' is inappropriate. The obvious example would be comparing the GPLMS model (South Africa) to the PRIMR model (Kenya). Both appear to be structured pedagogic interventions that combine different components, but the GPLMS's core training is two days twice a year and coaching, while PRIMR provided seven days of training and coaching by county education officials. Nadel and Pritchett (2016) made a strong case against large-scale RCTs and systematic reviews, and argue rather for a research approach they refer to as 'crawling the design space'. This approach focuses primarily on smaller scale piloting of multiple approaches with a stronger reliance on professional judgement to drive evaluation of what works.

Although the debate has shown the need to be sceptical about the policy value of the insights from the systematic reviews, we are unconvinced by the need to shift away from RCTs to 'crawling the design space' approaches. The EGRS experience suggests that there is a real need to understand how an intervention works at a scale that more or less mirrors the target population. Interventions that work for a handful of teachers may not work at scale. Studying change at scale is one of the great strengths of large RCTs. That said, the evaluation design of EGRS is strongly influenced by the Deaton and Cartwright argument about the importance of understanding mechanisms. But, understanding mechanisms is not just about economic theory and then validating the theory of change, but rather it is to investigate empirically how the actual intervention mechanism works at the coal face. It was therefore decided that the best way to understand the mechanisms is to complement the quantitative impact analysis (macro-level) with classroom observations in a sample of 60 schools (mezzo-level) and in-depth case studies (micro-level qualitative) and use social theory to make sense of data that emerges.

3. DESCRIPTION OF EGRS INTERVENTIONS

This study evaluates three different interventions, all aimed at improving reading and literacy in the home language, which in the case of the North West province is Setswana. The beneficiaries of the interventions were a cohort children entering Grade 1 at the start of 2015 over a two-year period (thus working with grade 2 learners in 2016). The project has now been extended for a third year of interventions at the grade 3 level in 2017. However, this extension does not apply to the parent involvement intervention.

Intervention 1: Structured lesson plans, additional reading materials + central training

Intervention 1 provides teachers with daily lesson plans, which are aligned to the curriculum as specified in the Curriculum and Assessment Policy Statements (CAPS) for home language literacy in the Foundation Phase. The four learning areas in the curriculum for grades 1 to 3 are Home Language literacy, First Additional Language (which is usually English), Mathematics, and Life Skills. The lesson plans are thus intended to strengthen the enactment of the curriculum and should not be seen as an alternative to current policy. They provide detailed specification for each lesson including information on methodology and content to be taught for each instructional day. The lesson plans incorporate the use of learning support materials including the government-provided "DBE workbooks" as well as certain additional materials (graded reading booklets, flash cards and posters), which are provided through the EGRS. The graded reading booklets provide a key resource for the teacher to use in group-guided reading and individual work so as to facilitate reading practice at an appropriate pace and sequence of progression. EGRS provided the Setswana "Vula Bula" graded reading book series developed by the Molteno Institute for Language and Literacy. These books were developed in the relevant African languages as opposed to being translated, and progress in accordance with the natural phonic progression of each language.

Intervention 1 trains the teachers on how to use the lesson plans and accompanying materials through central training sessions, each lasting 2 days, and occurring twice yearly. These sessions were conducted for grade 1 teachers in February and July of 2015 and for grade 2 teachers in January and July of 2016. Similar sessions are scheduled for 2017 as the project has been extended into grade 3.

Intervention 2: Reading Coaches, scripted lessons, graded readers.

Exactly the same set of instructional materials (structured lesson plans, graded reading booklets and other materials) is provided to Intervention 2 schools. Therefore, if the lesson plans are implemented with the same level of fidelity across Interventions 1 and 2, classroom practice and hence learning outcomes should be identical across the two groups. However, the modality of supporting teachers differs. Instead of bi-annual central training sessions, ongoing support to teachers consisting of regular (monthly) on-site coaching from specialist "reading coaches" is provided. In addition to these on-site visits, there are occasional meetings with the coach and a small cluster of nearby Intervention 2 schools. The evaluation of Interventions 1 and 2 should thus shed light on a) whether this structured lesson programme can improve the enactment of the curriculum and thus improve reading acquisition, and b) whether the mode of teacher support is important in determining effective enactment.

Intervention 3: Parental involvement

Intervention 3 is designed to promote parental involvement to support their children's reading progress. At each of the 50 schools in this Intervention group a Community Reading Coach (CRC) was recruited. The CRC was identified through communication with the school principal who recommended a suitably qualified and available person in the community. The CRCs attend a 1-day training session facilitated by the service provider (Class Act) at the start of each school term (quarterly). The CRCs are trained to deliver weekly training sessions for parents at their respective schools. For their services, CRCs are paid a stipend of R400 per month (about $35). Under this arrangement, CRCs are essentially volunteers receiving a small stipend, rather than employees receiving a salary.

A total of 30 sessions is scheduled for each year covering a total of 10 topics per year. Each topic has 3 sessions where the topic is the same but the activities of the session differ. Thus a parent can attend roughly 1 in 3 sessions and still be exposed to all topics, while parents who attend more regularly can still enjoy fresh activities. The topics covered in these sessions include the importance of learning to read for later educational and labour market success, training on how to support their child's reading at home and the provision of low-cost materials and reading games to use at home. As with Interventions 1 and 2, grade 1 parents were invited in 2015 and grade 2 parents in 2016.

4. SUMMARY OF CHANGES TO INTERVENTIONS SINCE THE START

Several minor changes have been made to the design of each of the three interventions since the programmes started being implemented. These alterations do not substantially affect the theory of change but are essentially designed to strengthen programme implementation. They also have minimal cost implications. Mid-way through 2015, the following changes were suggested by the implementation service provider and agreed to by the Research Team[5]:

- The establishment of Whatsapp groups amongst Intervention 1 teachers and trainers: Teachers in Intervention 1 attend central training once every six months. Therefore, it was felt that some channel for communication to those who provided the new materials and the training would be beneficial. Communication through Whatsapp groups was deemed to have virtually no cost implications and would be a suitable arrangement for any such support programmes involving centralized teacher training as a way to enhance sustained implementation.
- Start of term training for Intervention 2 teachers: It was decided that on-site coaching needed to be preceded by a light dose orientation training session at the start of every

[5] There were also several other changes suggested by the implementation service provider that were not agreed to by the Research Team.

term. This training does not occur at a single central venue but on several separate occasions and venues, hosted by each coach with their cluster of Intervention 2 schools. These orientation sessions last less than 1 full day.

- Symbolic rewards for Intervention 1 and 2 teachers: At training sessions, teachers are invited to present work done by their children and evidence of completing the prescribed learner assessments. Small non-monetary rewards are given to teachers who make successful presentations.
- Attendance incentives for parent attendance (Intervention 3): The big challenge experienced in Intervention 3 has been low levels of parent attendance. Therefore, we introduced a small cash incentive for attending the weekly parent meetings. Each week the Community Reading Coach conducts a lottery in which those parents in attendance compete for a prize of R25 (about $2). This is a small amount and the fact that only one parent can win it makes it a weak and partly symbolic incentive.

For 2016, several additional measures were agreed upon in order to improve parent attendance:

- Class Act were to ensure that Intervention 3 has a fixed routine – at some schools the meeting time was flexible and this was resulting in poor attendance rates;
- School principals have been involved in the management of the weekly meeting, monitoring the CRC and ensuring that they know what the training sessions entail;
- The CRCs were requested to explore other possibly convenient venues like the local churches in cases where transportation to the school is not available;
- A communication strategy using SMS messaging was implemented. Monthly SMS messages were sent to school principals to remind them to follow up on the CRCs and ensure that parent meetings occur. Class Act communicated with the CRCs on a weekly basis regarding what they were required to cover in that week;
- Principals were invited to the cluster CRC training that occurred once a month;

5. THEORY OF CHANGE

Reading acquisition

All three interventions relate to the educational theory of how reading acquisition occurs. An effective reader is one who reads with rich comprehension and engagement with the substance of the text. Reading comprehension is the product of two components: vocabulary and decoding.

To a great extent vocabulary (and language acquisition in general) comes naturally through hearing others speaking and then emulating this. Through speaking and hearing others speaking, phonological awareness also develops - this involves sound segmentation and recall of sound patterns. This phonological awareness is important for children to learn to decode

since written symbols are associated with particular sounds. Decoding thus consists of letter recognition and phonemic awareness.

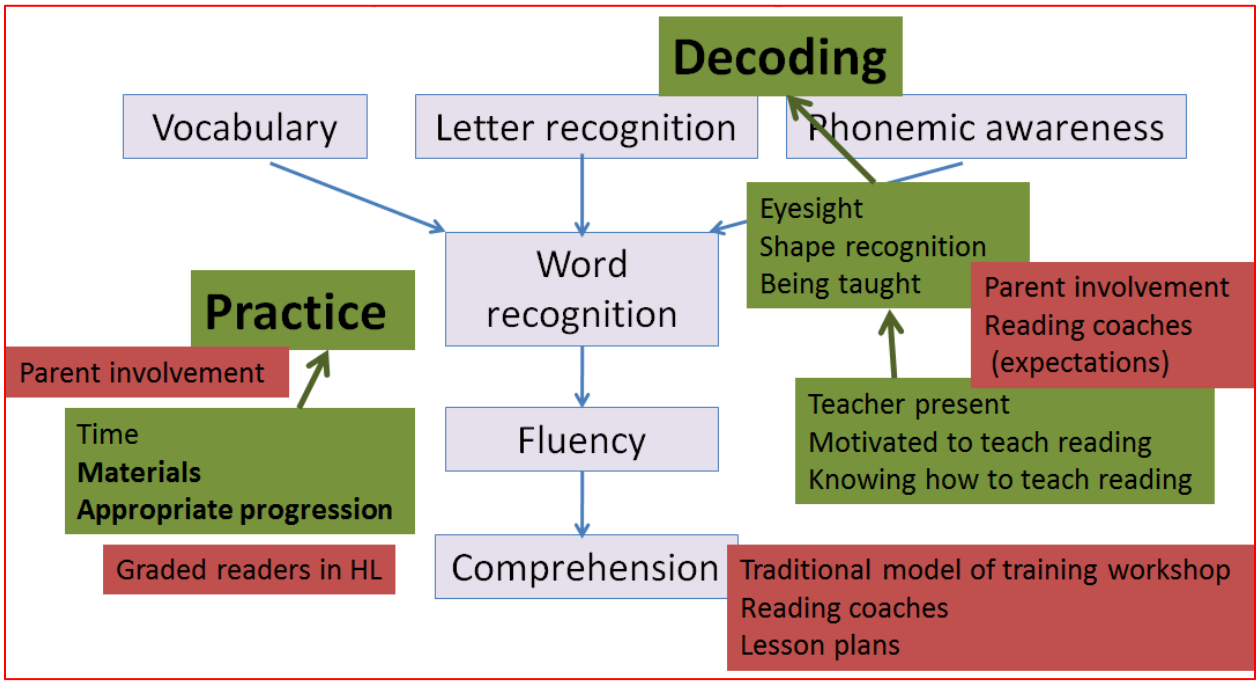
Unlike learning to speak, decoding does not come naturally; it is a method that must be taught systematically. It is important to emphasize that reading is produced by the product of vocabulary and decoding: If one has a perfect vocabulary but has not been taught the method of decoding one will not be able to read at all. Letter recognition and phonemic awareness are mastered through systematic teaching and consistent practice. This leads to the next stage of reading acquisition: word recognition. Through practice and appropriate progression from simpler sounds and words to more complex ones, word recognition becomes established leading to the next phase of reading acquisition: fluency. It is only once decoding and word recognition have become fluent, even to the point where it becomes automatic[6] and unconscious, that it is possible to reach the ultimate goal of reading comprehension. The strong empirical relationship between oral reading fluency and comprehension demonstrates this point.

> The ability to decode written symbols does not come naturally – it must be systematically taught.

In order to learn the basics of decoding, a child requires a teacher who is present, capable and motivated to deliver systematic reading instruction. In order for decoding to become fluent a child requires suitable graded materials and the discipline (perhaps imposed) to practice a lot. The interventions to be tested in this study address these needs in various ways. Figure 1 presents a theoretical diagram illustrating how reading acquisition occurs, what supportive conditions need to be in place and how each of the interventions being evaluated in the EGRS address key stages in the development of reading acquisition.

[6] For example, do not read this word: “exactly”.

Figure 1: Theoretical diagram of how reading acquisition occurs



There is a growing body of evidence from developing countries that early grade reading interventions can have a significant impact. The "EGRA Plus" programme administered in Liberia, for example, produced substantial gains in reading achievement relative to comparison children who did not receive the programme. Key aspects of this programme included a cascading model of reading coaches, the distribution of scripted lesson plans and reading assessment tools, and the dissemination of report cards to parents (Gove and Wetterberg, 2011).

A richer discussion of international studies using structured learning programmes and coaching will be provided in the literature review section. However, these studies often do not tell us which component of the intervention was responsible for the success of the program. This is important for policy purposes, because we want to find the most cost-effective intervention which could be scaled up by government. For example, the "EGRA plus" programme in Liberia was clearly highly resource-intensive because it required ongoing monitoring from qualified reading coaches, but we do not know if one might be able to reach the same results with a sub-component of the program. Moreover, there is uncertainty about the transferability of the findings given different language and social contexts.

Similar programs have been implemented in South Africa, but since they were not credibly evaluated, we do not know if they truly improved pupils' reading acquisition. The Department of Basic Education typically holds training programs similar to our intervention 1; and Gauteng has implemented a model of reading coaches, similar to intervention 2. Since it has not been possible to produce a robust empirical impact evaluation of these programmes, we have little

knowledge about their effectiveness. Fleisch and Schoer (2014) attempted a Regression Discontinuity Design (RDD) to evaluate the impact of the Gauteng Primary Language and Mathematics Strategy (GPLMS) and findings pointed to a positive impact, though the findings were tentatively made given significant data constraints. Sailors et al (2010) evaluated a reading intervention in South Africa, which followed a similar model to intervention 2, but there are considerable methodological limitations to this study. These South African studies are discussed more extensively in the literature review.

There have also been several experiments around the world that have provided information to parents with the goal of fostering parental involvement and thus improving learning outcomes. But there is much we still do not know. In Pakistan, pupils who came from villages where the community was provided with information of school performance performed better in independently administered tests, compared to pupils from villages where no such information was administered. The improvement was particularly large for schools with low initial learning outcomes (Andrabi, Das, & Khwaja, 2015). In a different programme in India, school communities were informed of their school performance and also educated on their rights, roles and responsibilities in school governance through 8 public meetings. Education performance improved as a result (Pandey, Goyal, & Sundararaman, 2009). However, in a recent impact evaluation in Kenya, informing parents on their child's reading progress had zero impact (Lieberman, Posner and Tsai, 2013). The authors hypothesize necessary conditions for an information-intervention to work, all of which we address in our study: (i) information is new; (ii) it highlights under-performance and potential to improve; (iii) it is combined with measures which enable parents to act on this information.

All interventions aim to improve reading acquisition in the home language. Strictly speaking, the targeted outcome is home language literacy more broadly, since this is the Foundation Phase curriculum area being given support through our programmes. The choice to address home language literacy is motivated by research showing long-term benefits to strong home language skills prior to switching to a second language. Taylor and Von Fintel (2016), for instance, show that in South Africa using home language as the language of instruction during grades 1, 2 and 3 has been associated with better English acquisition in grades 4, 5 and 6.

Intervention 1:

There are a number of reasons why one might expect a structured learning programme using scripted lesson plans to improve instructional practice and learning. Firstly, we know that curriculum coverage, pacing and sequencing is currently inadequate in the majority of South African classrooms. The National School Effectiveness Study revealed this through a learner exercise book review conducted in a large sample survey (Taylor, Van der Berg, & Mabogoane, 2013) while classroom observation studies have unpacked this in greater detail (Hoadley, 2010). A structured learning programme clearly has the potential to improve curriculum coverage, pacing and sequencing.

Secondly, the use of lesson plans can facilitate the adoption of new methods by teachers and thus expand their own repertoire of instructional practices. Most teacher training interventions implicitly assume that changed knowledge will lead to changed practice in the classroom. However, the use of lesson plans allows one not to rely completely on this assumption. When a lesson plan prescribes the use of a certain instructional method, the teacher may implement that method even though she may not yet possess a deep understanding of the rationale behind the method. Through the regular practice of that method, however, the teacher's knowledge may be enriched as they begin to see the method's effectiveness. In this way there is an iterative relationship between "knowing" and "doing" in which improved classroom practice emerges.

Lesson plans facilitate the adoption of new instructional methods and increase the teacher's instructional repertoire.

A third reason to expect lesson plans to improve classroom practice is that they integrate the effective use of reading materials. Van der Berg (2008) made the case that additional school resources often make no impact because they are not well managed by schools. An HSRC study of grades 1 – 4 classrooms in 20 Limpopo schools found that little reading activity occurred, that the use of texts was limited and that when reading was taught the predominant activity was the teacher reading to the class (Prinsloo, 2008). One of the most important national interventions over the last few years has been the provision of the so-called "DBE Workbooks". These colourful books are a type of hybrid between a textbook and an exercise book, with lots of exercises for learners to complete in the books themselves. The lesson plans provided through EGRS incorporate the DBE Workbooks into the daily lessons referring to specific page numbers for exercises to complete. The additional reading materials provided through EGRS (posters, flash cards and graded reading books) are similarly integrated into the structured learning programme through the daily lesson plans. Several reports bemoan a lack of African language reading materials in Foundation Phase classrooms (NEEDU Report, 2012; Ministerial Audit of provincial reading programmes, 2012). Learning to read requires practicing and gradually moving from simple language structures to more complex letter blends and words. The language structures differ across languages and translations of reading booklets is therefore inappropriate. Interventions 1 and 2 in the EGRS fill this gap by providing sets of Vula Bula graded reading booklets developed in Setswana, and by promoting their effective use through the lesson plans.

A structured learning programme with daily lesson plans can ensure that reading resources are effectively used.

Yet there are potential negative (or perhaps ambiguous) consequences of following a prescriptive set of lesson plans. As is the case in many developing countries, South African classrooms often comprise a wide range of learner proficiency levels. For example, in our control group sample of 80 schools, nearly 40% of grade 2 children could not read a single word

in the paragraph reading test; yet about 25% of children could read at least 50 words in a minute. And this sample already excludes the two most affluent "quintiles" of schools, making it a *relatively* homogeneous sub-set of South African schools. Some might argue that scripted lesson plans could reduce teacher autonomy to differentiate the level of instruction to meet the variety of needs present within the classroom. If this is the case, the structured learning programme may benefit a certain range of the learner proficiency distribution depending on where the lessons are pitched. This is definitely something that we are aware of and will test for. However, there is one aspect of the EGRS learning programme that should in fact promote differentiated instruction, namely that the lessons routinely use "group-guided reading" sessions. This activity, which is prescribed in the CAPS, involves a set of between 6 and 10 learners sitting with the teacher to read selected reading material. This activity promotes individualized attention to learners and thus promotes the opportunity for individual decoding as opposed to whole class reading or "chorusing" after the teacher. Also, the programme encourages teachers to group learners according to their level of proficiency, thus promoting a degree of differentiated learning.

Intervention 2:

The reading coach intervention provides basic orientation to the lesson plans and additional reading materials at the start of each term followed up by on-site coaching visits approximately once a month. The fact that the coach actually observes classroom practice makes it more likely that teachers will in fact implement the new practices as prescribed in the lesson plans. Moreover, it promotes the correct implementation of instructional methods since the coach is able to indicate things could be done differently. In this way, the coach can guide a teacher through the process of trying a particular method, reflecting on that activity, and doing it again in an improved form. Thus, the iterative relationship between "knowing" and "doing" is strengthened in Intervention 2 by the presence of a coach who acts as a mentor along the way.

The assumption is that, just like learning to read, the ability to teach is a skill that needs to be developed over time and might not be accomplished in one-off training. Furthermore, the reading coaches could also improve teacher motivation as they are frequently monitored, provided with much-needed additional support, and can also find inspiration from watching an excellent example provided occasionally by coaches. This programme thus addresses both teacher capacity and teacher motivation. Another way to describe the difference between Interventions 1 and 2 is that while they share an underlying pedagogical theory of change (centered around instructional alignment and coherence using prescriptiveness as a vehicle), they differ in their theory of action (where Intervention 2 has a stronger component focused on changing behavior using accountability and motivation).

Intervention 3:

Parents pay a critical component to learning to read, as it requires continuous practice, both at school and at home. For parents to be *willing* to play this role they need to appreciate (i) the importance of reading; and (ii) that their child is most likely not learning enough at school and requires additional support. This is the purpose of the information. For parents to be *able* to play this role, they need to understand the necessary steps in learning to read and also have appropriate material to practice reading with their child. This is the purpose of the training and additional practice material.

6. RESEARCH SITE

The EGRS is being implemented in the North West province, in the districts of Dr Kenneth Kaunda and Ngaka Modiri Molema. The North West province was chosen on the basis of 1) it being a relatively poor province, thus making it relevant to the majority of the underperforming South African school system; 2) it is relatively homogenous in terms of home language (Setswana) making it more affordable to develop learning support materials in a single language; 3) it is within driving distance from the Gauteng province where the national DBE is located; and 4) the senior management of the North West provincial education department were eager to partner with the DBE on this project. The district of Bojanala was excluded because another special targeted intervention was taking place in that district at the same time. The district of Dr Ruth Segomotsi Mompati was excluded since it is particularly far West of Gauteng and since enough schools existed in the districts of Dr Kenneth Kaunda and Ngaka Modiri Molema. Figure 2 shows a map of South Africa divided into the 83 education districts.

Figure 2: Map of South Africa showing education districts

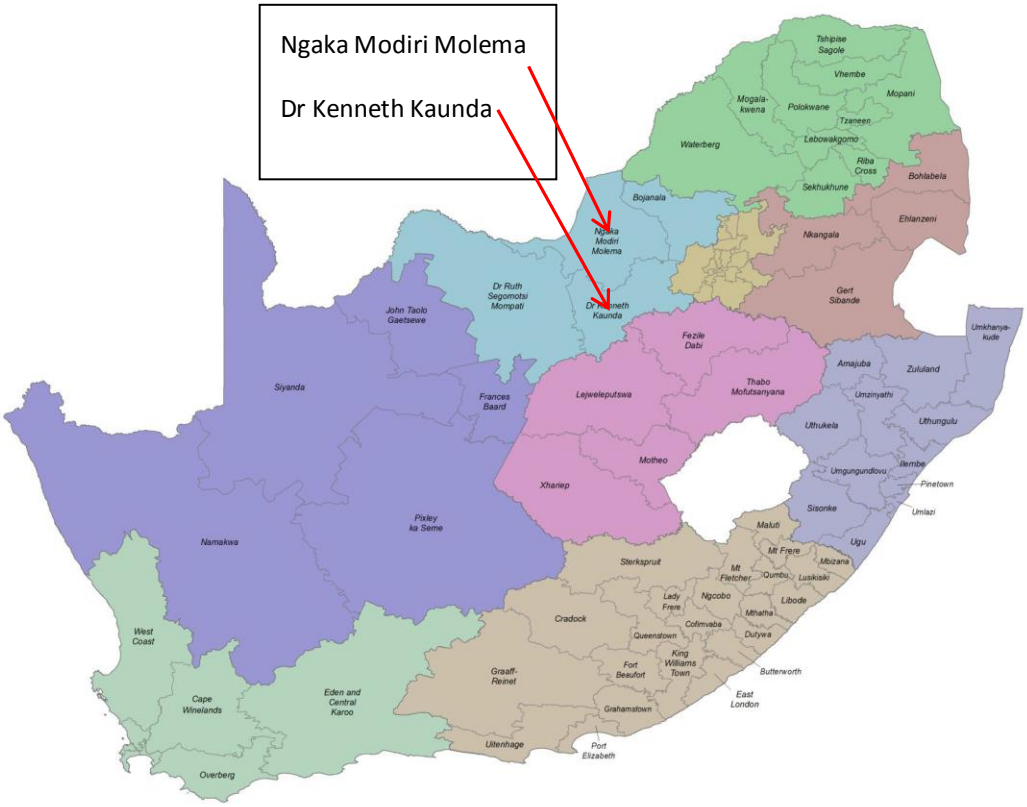


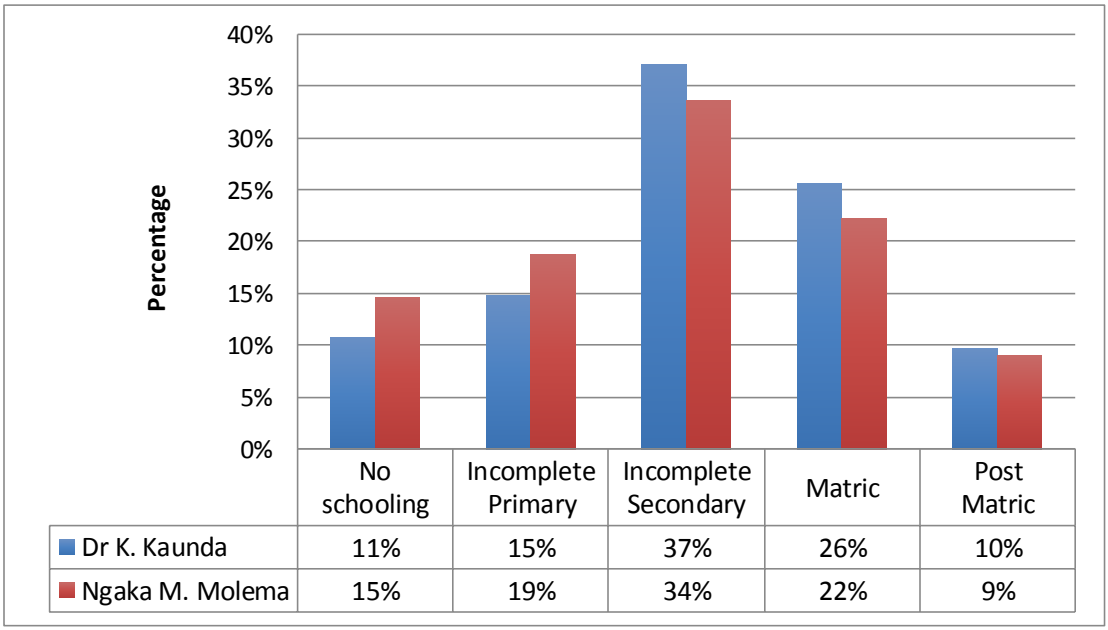
Table 1 below shows the total number of ordinary schools by phase for both Dr Kenneth Kaunda and Ngaka Modiri Molema districts in 2014. We see that Ngaka Modiri Molema district has the highest number of schools across all categories. Of the 248 schools in Dr Kenneth Kaunda district, 14 are independent schools while 11 of the 404 schools in Ngaka Modiri Molema district are independent schools. In Dr Kenneth Kaunda, 81% of schools are no-fee schools (classified as Quintile 1, 2, and 3 according to the official school poverty classification) while the equivalent figure was 91% of schools in Ngaka Modiri Molema district. This confirms that these two districts are indeed poor and rural parts of South Africa. The choice of these areas for the EGRS project was deliberate so as to optimize the relevance of the study's findings to the large, underperforming and poor sections of South Africa's school system.

Table 1: Number of schools by phase in Dr Kenneth Kaunda and Ngaka Modiri Molema

	Dr Kenneth Kaunda		Ngaka Modiri Molema	
	Number	%	Number	%
Primary	149	60%	247	61%
Secondary	54	22%	76	19%
Combined	42	17%	67	17%
Intermediate	3	1%	14	3%
Total	248	100%	404	100%

In the 2011 Census, people were asked to indicate the highest level of education that they had completed. It referred to the highest level completed, not the level currently in, if the person was still studying. Figure 3 shows the education levels of adults aged 20 and older by district. The category 'Matric' refers to the secondary school leaving examination. This figure shows that Dr Kenneth Kaunda district had slightly higher proportions of people with matric and post matric qualifications compared to those in Ngaka Modiri Molema district. Overall, this figure implies that the majority of people who would be parents to Grade 1 and 2 pupils would have relatively low levels of education.

Figure 3: Highest Education level for adults aged 20 and older



	No schooling	Incomplete Primary	Incomplete Secondary	Matric	Post Matric
Dr K. Kaunda	11%	15%	37%	26%	10%
Ngaka M. Molema	15%	19%	34%	22%	9%

Source: Census 2011

The Annual National Assessments (ANA) provide an indication of school performance at the primary school level.[7] It should be noted, however, that results are not comparable across time or across subjects or grades, since the tests cannot be equated to each other. In 2012 Dr Kenneth Kaunda performed better than Ngaka Modiri Molema. However, the opposite was true in 2013. This seems strange, and may reflect differential test administration and marking practices across time and district. The broad point to note is that language and mathematics performance in both of these districts is at a low level, allowing much room for improvement.

[7] The last time the ANA was implemented was in 2014. The national assessments system is currently being redesigned.

Table 2: Grade 3 learners achieving 50% and above by subject

Subject	Year	Dr Kenneth Kaunda	Ngaka Modiri Molema
Mathematics	2012	30%	18%
	2013	49%	48%
Language	2012	53%	41%
	2013	44%	49%

Table 3: Grade 6 learners achieving 50% and above by subject

Subject	Year	Dr Kenneth Kaunda	Ngaka Modiri Molema
Mathematics	2012	9%	7%
	2013	15%	23%
Language	2012	25%	19%
	2013	40%	45%

7. EVALUATION DESIGN

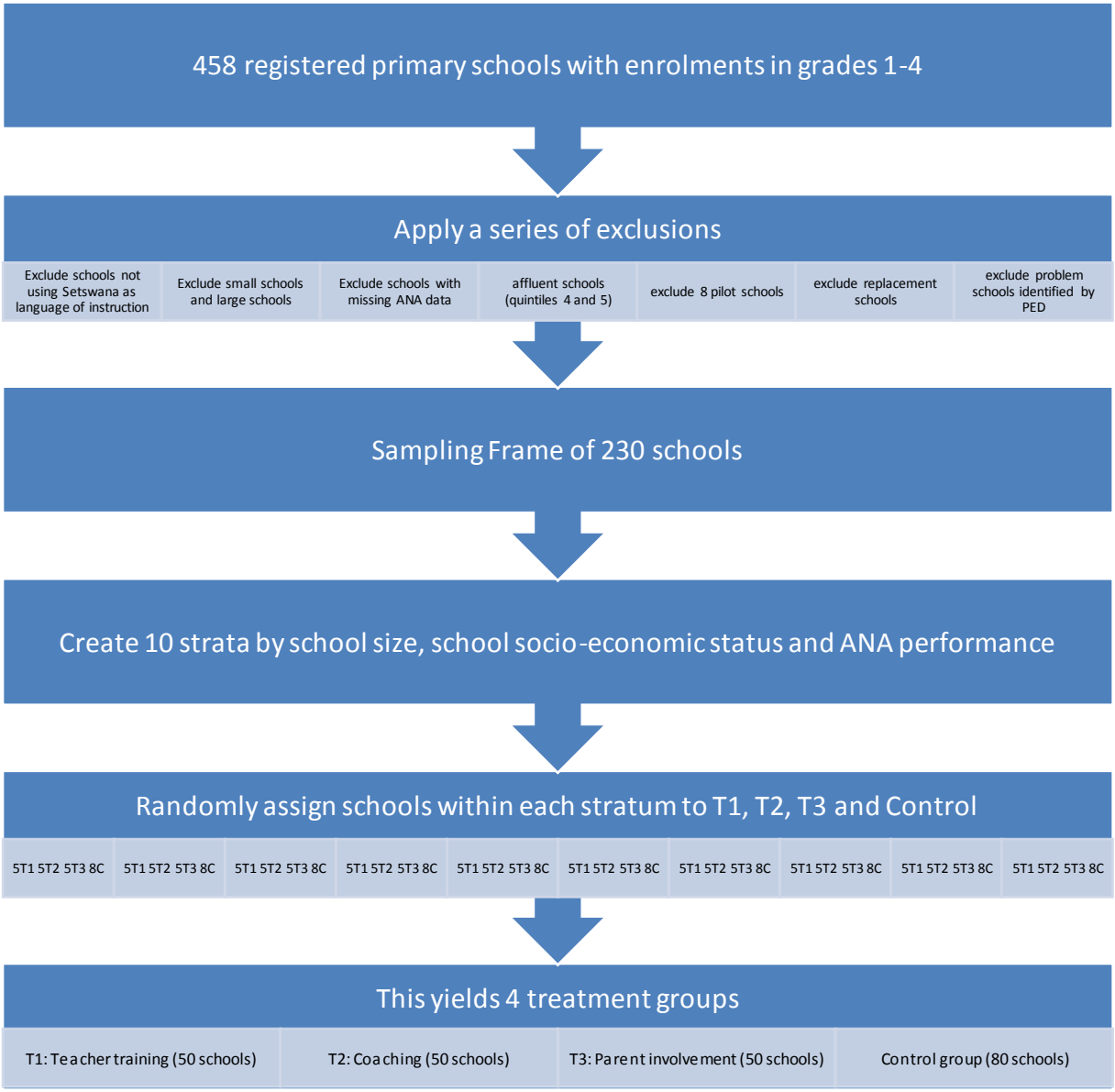
SAMPLE SELECTION AND ASSIGNMENT TO INTERVENTION GROUP

Through a process of elimination we developed a sampling frame of 230 eligible schools. Beginning with 458 primary schools in the districts of Dr Kenneth Kaunda and Ngaka Modiri Molema registered according to 2014 administrative data, we started by excluding relatively affluent schools (those in quintiles 4 and 5). Next, we excluded schools in which the language of instruction in the Foundation Phase was not Setswana. We then excluded schools which were missing in the 2014 ANA dataset. We also excluded 8 schools that had already been selected for the purposes of piloting of instruments through the course of this project. We further excluded particularly small schools (fewer than 20 grade 1 enrolments) since many of these schools would practice multi-grade teaching rendering the grade-specific lesson plans less appropriate. We also excluded particularly large schools (more than 180 grade 1 enrolments) to limit intervention costs. Three more schools were excluded after the North West PED checked our list of schools and found specific problems with these schools (e.g. the school had been closed down, or a particular conflict around school management was occurring). After all of these exclusions 235 eligible schools remained. Using a random number generator, we then excluded 5 schools, which we retained as possible replacement schools. Thus we obtained the sampling frame of 230 schools.

To increase power and balance between Intervention arms, we performed stratified randomization. We created 10 strata of 23 similar schools based on school size, socio-economic status, and performance in the Annual National Assessments. Within each stratum,

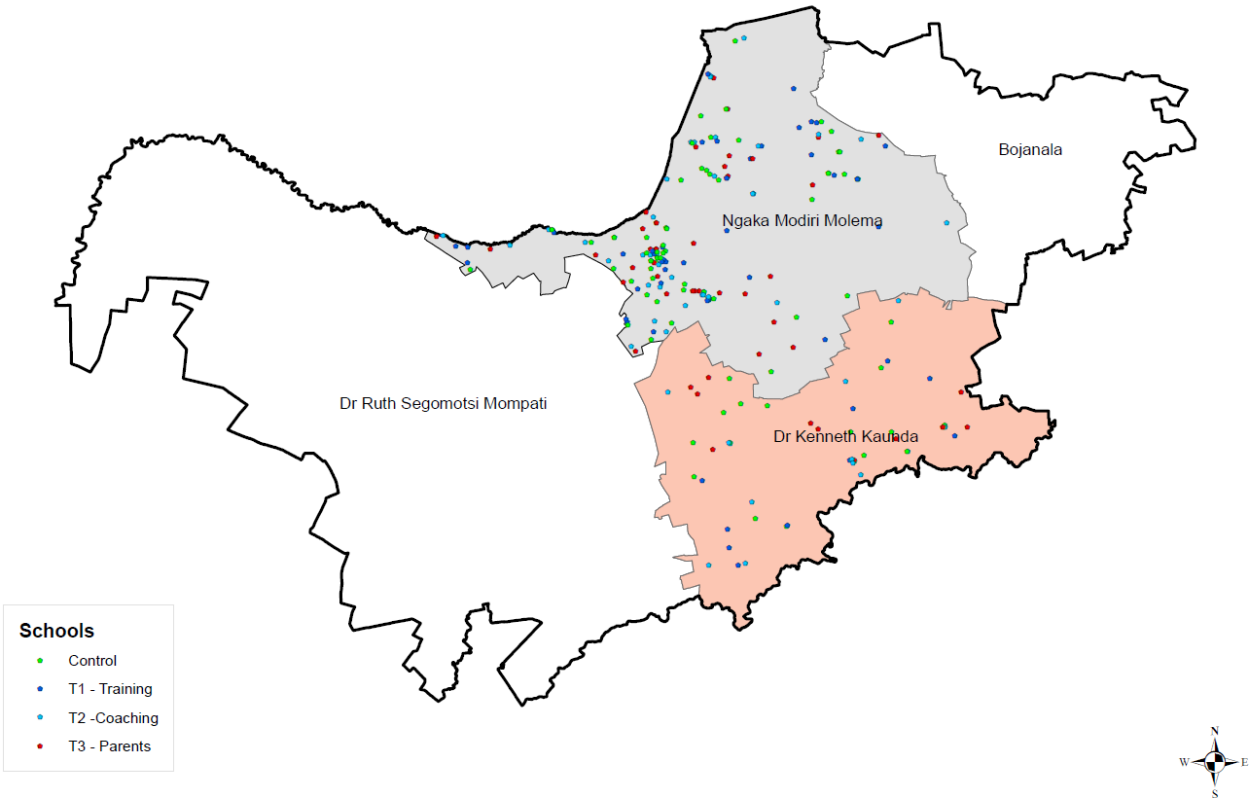
we then randomly assigned 5 schools to each Intervention group and 8 to the control group. Thus we randomly assigned 50 schools to each Intervention and 80 to the control. Given that we collect data on 20 grade 1 learners per school, this sample should be sufficient to identify a minimum detectable effect size of 0.21 standard deviations when comparing an Intervention group with the control group and a minimum detectable effect size of 0.23 standard deviations when comparing two Intervention groups. These calculations assume a 95% confidence interval, an alpha value of 0.8, an intra-class correlation coefficient (rho) of 0.3 and a correlation between pre- and post-test scores of 0.7. Figure 4 presents a schematic diagram to describe the sampling procedure that was followed.

Figure 4: Diagram showing sampling procedure



The following map shows the schools participating in the EGRS and indicates the Intervention status of each school. Note that a few schools are not shown on the map due to missing or inaccurate GIS codes.

Figure 5: Map of North West province showing schools by Intervention assignment



8. SUMMARY OF BASELINE RESULTS

This section summarizes the results of the baseline data collection, which was conducted at the start of grade 1 in 2015. More details are available in the full Baseline Report.

DATA COLLECTION AND CAPTURING

As described in the full Baseline Report, the random selection of 20 learners per school appears to have been effectively done, and a high proportion of the intended sample of learners was successfully tested at baseline. However, there were several problems noted with regard to the quality of fieldwork resulting in low response rates to parent, teacher and principal questionnaires.

LEARNER TEST STATISTICS

The main objective with baseline learner assessment was to obtain measures of learner ability/proficiency that would be predictive of later reading outcomes. This is because the statistical power of the impact evaluation is positively related to the correlation between baseline and outcome measures. Put differently, the more one is able to account for prior influences on reading outcomes, the more precisely one can identify the impact of the interventions.

For this reason, the baseline test was not curriculum-specific or standard-based, but was designed to include predictive measures of learning to read, such as the "digit span" test of working memory. This test item involves the test administrator speaking two unrelated words and the learner then being required to repeat them from memory. The process is then repeated with three words, then four words, five and ultimately six words. The same procedure is then followed using numbers rather than words. This digit span exercise is of course not a reading outcome, but it is believed to be predictive of the capacity to decode and thus learning to read.

It was also important to avoid so-called "floor" and "ceiling" effects. A floor effect is when a substantial proportion of learners score zero on a test. The problem with this is that there is in fact variation in ability amongst those scoring zero but that variation is occurring beneath the "floor" of the test and is thus unobserved. The test thus has no predictive power amongst those learners. Conversely, a ceiling effect occurs when a substantial proportion of learners score the maximum score possible. The test thus has no ability to distinguish amongst the more proficient learners. In order to avoid floor effects the picture comprehension or expressive vocabulary section was included since this skill is likely to be widely developed even amongst learners with no ability to decode letter symbols. Conversely, harder items which one would not necessarily expect a learner entering grade 1 to be able to do, such as sentence reading, were included in order to avoid a ceiling effect on the test.

Table 4 reports the summary statistics for the Baseline sub-tests as well as a composite test score derived using Principal Components Analysis (PCA)[8]. The composite score was adjusted so that it would have a mean of zero and a standard deviation of 1. Sub-test A (picture comprehension) indeed turned out to be the easiest section, with more than 25% of the sample getting 10 out of 10. The one sub-test where there was neither a floor nor a ceiling effect was Section C (Working Memory). Less than 10% of children scored at below 2 out of 10 (indicating

[8] In calculating a composite score one needs to decide how much weight to attach to each subtask in the test. One cannot calculate simply add each subtask's score together, since one subtask may have had more items but should not necessarily carry more significance than another subtask. Therefore, we ran Principal Components Analysis (PCA) on the subtotals for each subtask. In PCA the variation within all variables included is analysed and those linear combinations capturing the most common variation amongst variables are identified. It is assumed that the linear combination, referred to as a principal component, which captures the most common variation amongst the variables included represents the underlying construct of interest. In this case we might think of the primary underlying construct being measured as reading ability. The weight given to each variable when calculating the total composite score is then determined by the extent of that variable's correlation with the first principal component. The intuition is that a subtask that is not well correlated with the other subtasks may be measuring something different from the intended underlying construct – this subtask should therefore carry less weight in a composite index.

no floor effect) while less than 10% of children scored more than 8 out of 10 (indicating no ceiling effect). The other sub-tests all had floor effects – with at least 25% of children scoring zero. However, since Sub-tests A and C did not have floor effects the overall test did not have a strict floor effect.

Table 4: Summary statistics – Baseline sub-tests

	count	mean	min	p10	p25	p50	p75	p90	max
A. Picture Comprehension	4538	8.58	0	7	8	9	10	10	10
B. Letter sound recognition	4538	5.08	0	0	0	2	6	13	99
C. Digit span (Working memory)	4538	4.99	0	2	4	5	6	8	10
D. Phonological awareness	4538	2.17	0	0	0	1	3	7	12
E. Word recognition	4538	1.9	0	0	0	0	2	5	50
F1. Sentence reading	4538	0.73	0	0	0	0	1	3	3
F2. Sentence comprehension	4538	1.22	0	0	0	0	0	3	15
Y1. Composite Baseline Score	4538	0	-1.83	-0.82	-0.58	-0.29	0.27	1.13	5.40

Note: "p10" refers to the score at the 10th percentile of the distribution, "p25" to the score at the 25th percentile, and so forth.

One way to get an indication of the validity of the baseline sub-tests is to consider the correlations between sub-tests, as reported in Table 5. Overall, the correlation coefficients are somewhat on the low side, certainly compared to the intra-test correlations observed at Midline and Endline, as will be reported later. If we consider the predictive power of the baseline sub-tests by observing the correlations of each sub-test with the Midline composite score and the Endline composite score it is disappointing how low the correlations are. In contrast, the correlation between the Midline composite score and the Endline composite score is high at 0.72. Therefore, the Baseline is the "noisy" measure not well correlated with the rest. There are two main reasons why we think the Baseline provided a rather noisy measure. Firstly, the Research Team had concerns about the quality of work done by the fieldwork service provider, which was subcontracted by the HSRC. In particular, the recruitment procedures were not satisfactory, nor were the logistical arrangements regarding the fieldwork schedule and transportation of fieldworkers to schools. After various quality assurance measures and a new service provider for the Midline and Endline there was a marked improvement in the quality of work, and this is borne out in the data. Secondly, the younger children are the more difficult it is to design good test instruments that are nevertheless conducive to standardized administration by non-specialist fieldworkers. For higher grades, one can simply provide a pen and paper test to children and all the fieldworker has to do is distribute and collect the instruments. But with one-on-one testing with young children it is a lot harder to ensure that fieldworkers administer and interpret the testing correctly and consistently.

A further point to note from Table 5 is that the letter recognition sub-test was the highest correlated baseline sub-test with the Midline and the Endline scores. This was followed by the working memory and phonological awareness sub-tests. In fact, letter recognition was more highly correlated with Midline and Endline scores than the composite baseline score. Partly for

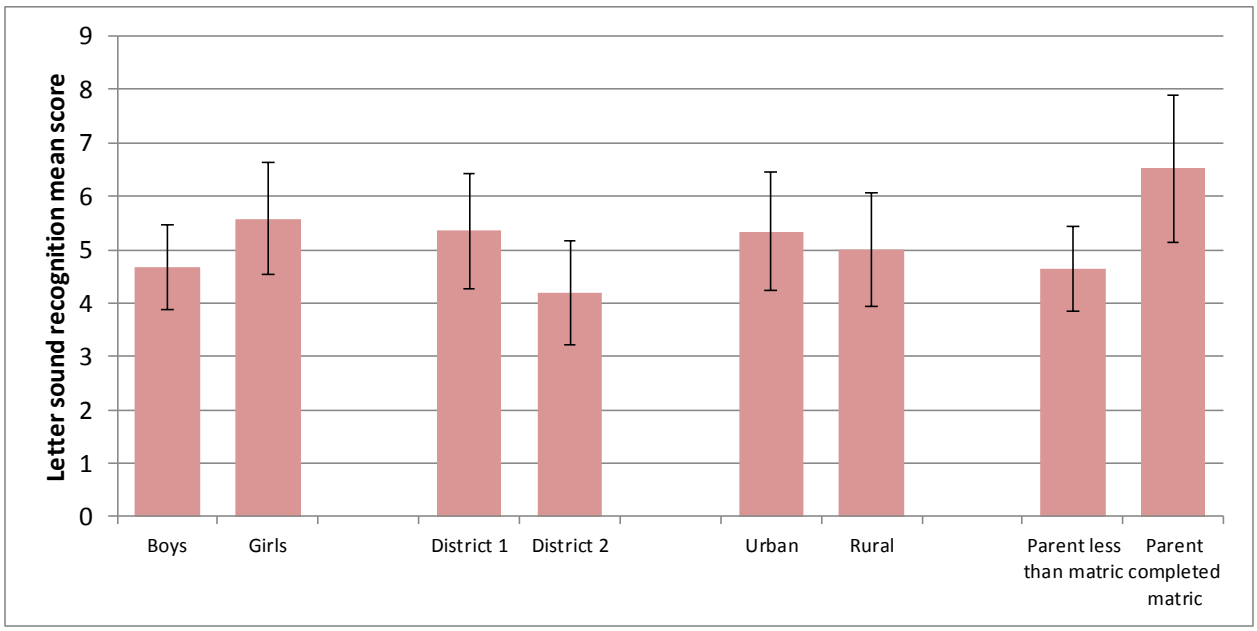
this reason, the impact evaluation model specifications used later on control for all baseline sub-test scores individually, rather than through the inclusion of the composite baseline score.

Table 5: Correlation matrix of Baseline sub-tests, Wave 2 and Wave 3 composite scores

	[A]	[B]	[C]	[D]	[E]	[F1]	[F2]	[Y1]	[Y2]	[Y3]
A. Picture Comprehension	1									
B. Letter sound recognition	0.14	1								
C. Digit span (Working memory)	0.30	0.24	1							
D. Phonological awareness	0.23	0.34	0.46	1						
E. Word recognition	0.17	0.53	0.33	0.47	1					
F1. Sentence reading	0.15	0.18	0.22	0.37	0.35	1				
F2. Sentence comprehension	0.14	0.23	0.31	0.45	0.46	0.53	1			
Y1. Composite Baseline Score	0.39	0.58	0.62	0.76	0.76	0.63	0.72	1		
Y2. Wave 2 composite score	0.12	0.34	0.18	0.19	0.18	0.10	0.05	0.25	1	
Y3. Wave 3 composite score	0.11	0.26	0.19	0.18	0.13	0.09	0.06	0.22	0.72	1

Figure 6 presents the mean letter recognition scores at baseline for various sub-groups of interest. On this item, as with others, girls performed better than boys at the start of school. There may simply be physiological reasons for the difference in achievement by gender at this age. However, it will be interesting to track the gap between boys and girls because we know from numerous other assessments in higher grades that in South Africa girls are performing better than boys in literacy (and even to some extent in mathematics), are less likely to repeat grades than boys, and are more likely to complete secondary education than boys. As will be shown later in this report, the gap between girls and boys is actually larger at Midline and at Endline than at Baseline, but this could be partly due to the noisy nature of the Baseline data. Figure 6 also shows a difference in baseline achievement between the two education districts represented in the sample. There is no real difference between children attending schools in urban and rural areas. Children in households where they have at least one parent or guardian with at least complete secondary education achieved noticeably higher than children in homes with less parental education.

Figure 6: Mean scores by sub-groups at Baseline



Note: 95% Confidence intervals are shown

BALANCE TESTS

In the Baseline Report we indicated that there was near universal balance across treatment arms. However, we have subsequently discovered an error in the data analysis, and we now are finding that Treatment 1 (Training) had achieved statistically significantly lower scores on several of the baseline sub-tests. This is strange given that the random assignment was carried out with fidelity. Some analysts recommend not reporting baseline balance since randomization allows one to assume balance and it is always possible that a degree of imbalance might exist. Nevertheless, we feel that since we reported an error it is important to publish the corrected numbers here. Table 6 shows the results of regressions to test if the differences in average scores in learning outcomes between treatment groups are statistically significantly different from zero. Each column shows a separate regression on treatment indicators after controlling for strata fixed effects. The standard errors are clustered at the school level. One star indicates that the difference in means between one of the treatments and the control is statistically significant at the 10% level. The bottom three rows show the p value for the equality tests on the treatment coefficients. In other words, it shows the pair-wise tests comparing the means between treatment groups. Out of the 42 possible comparisons, there is slight imbalance in 6 cases, all involving Treatment 1. This slight imbalance should not bias our main conclusions since we do control for baseline learner scores in our main model specifications.

Table 6: Baseline balance tests

	(1) Expressive vocabulary	(2) Letter Recognition	(3) Working memory	(4) Phonological awareness	(5) Word recognition	(6) Sentence comprehension	(7) Words in sentence	(8) Combined score
T1 (training)	-0.272*	-1.261	-0.491*	-0.789**	-0.680	-0.0673	-0.0609	-0.209*
	(0.149)	(1.101)	(0.283)	(0.348)	(0.542)	(0.188)	(0.359)	(0.119)
T2 (coaching)	-0.0824	0.419	-0.156	-0.193	0.199	0.248	0.597	0.0670
	(0.138)	(1.215)	(0.296)	(0.406)	(0.715)	(0.195)	(0.509)	(0.147)
T3 (parents)	-0.215	-0.652	-0.282	-0.290	0.0898	-0.115	0.810	-0.0389
	(0.140)	(1.209)	(0.308)	(0.405)	(0.735)	(0.171)	(0.553)	(0.153)
Obs	4538	4538	4538	4538	4538	4538	4538	4538
T1=T2: p-value	0.233	0.106	0.289	0.154	0.216	0.155	0.203	0.070
T2=T3: p-value	0.721	0.557	0.524	0.231	0.294	0.811	0.120	0.282
T1=T3: p-value	0.381	0.354	0.710	0.834	0.900	0.080	0.749	0.555
Control mean	8.704	5.406	5.196	2.450	1.994	0.719	0.926	0.040

*** p<0.01, ** p<0.05, * p<0.1
Standard errors in parentheses

9. SUMMARY OF MIDLINE RESULTS

As described in the full Midline Report, a number of measures were put in place by the HSRC and the Research Team to ensure a better quality of fieldwork in the Midline survey (October/November 2015) compared to what had occurred in the baseline data collection. The Terms of Reference for the subcontracting of a fieldwork agency was much more detailed with respect to fieldworker selection criteria, conditions around approval of and payment for deliverables, and overall functionality criteria for the fieldwork organization. Instead of a single day of fieldworker training, there was a three-day training programme for fieldworkers including a practice round of data collection (with monitoring and feedback) at five schools not included in the project. The Terms of Reference specified that exactly 40 fieldworkers should be recruited, 20 of whom will administer the learner tests and must have expertise in early grade teaching. The fieldwork schedule needed to be submitted well in advance to the HSRC with schools already having been contacted and appointments fixed for specific days made. As documented in the Midline Report, this led to a different fieldwork service provider being appointed at Midline and the quality of work was considerably better. Communication with schools ahead of the fieldwork visit was also much smoother since we had an updated database of contact information, which the DBE compiled using information collected in baseline questionnaires and by the implementing agent for interventions. Finally, extensive revisions were made to the midline instruments, especially the shortening of the school principal and teacher questionnaires, with the intention of improving response rates.

LEARNER TEST SCORES

The Midline learner assessment instrument was adapted from the Setswana Early Grade Reading Assessment (EGRA), and was similar to the Baseline assessment instrument. Three of the sub-tests in the Midline assessment (Letter Recognition, Word Recognition and Phonological Awareness[9]) remained exactly the same as in the Baseline assessment and can therefore be used to directly assess learning gains made over the year. The summary statistics for the sub-tests in the Midline learner assessment are presented in Table 7. Floor effects were observed to a greater extent than we expected based on the midline pilot. Only the writing sub-test did not have a floor effect. For non-word decoding, paragraph reading, comprehension and phonological awareness we observe scores of zero at the median. Fortunately, however, for letter recognition and writing there was a fair amount of variation within the bottom 25% of learners. There were definitely no ceiling effects (which would have occurred had the test lacked in difficult items).

Table 7: Summary statistics for each sub-test in Midline learner assessment

Sub-test	min	p10	p25	p50	p75	p90	max
Letter recog.	0	0	4	16	38	54	110
Word recog.	0	0	0	3	9	22	50
Non-word decoding	0	0	0	0	6	18	50
Sentence reading	0	0	0	1	9	11	11
Paragraph reading	0	0	0	0	11	30	64
Comprehension	0	0	0	0	1	4	6
Writing	0	1	4	6	8	11	12
Phonological awareness	0	0	0	0	1	3	4
Combined score	-0.943	-0.868	-0.718	-0.444	0.486	1.693	3.650

As depicted in Table 7, the correlation coefficients between the midline subtests were considerably higher than those within the baseline test. This may reflect the better quality fieldwork at midline compared with the baseline, since one would expect the quality of fieldwork (consistency in how items were administered and scored across fieldworkers and learners) to be directly related to how accurate a signal the tests provide of learner proficiency. With completely haphazard fieldwork one would expect zero correlation across sub-tests. This is confirmed by the fact that the midline subtests were all more strongly correlated with the Wave 3 composite score than they were with the baseline composite score (Table 8).

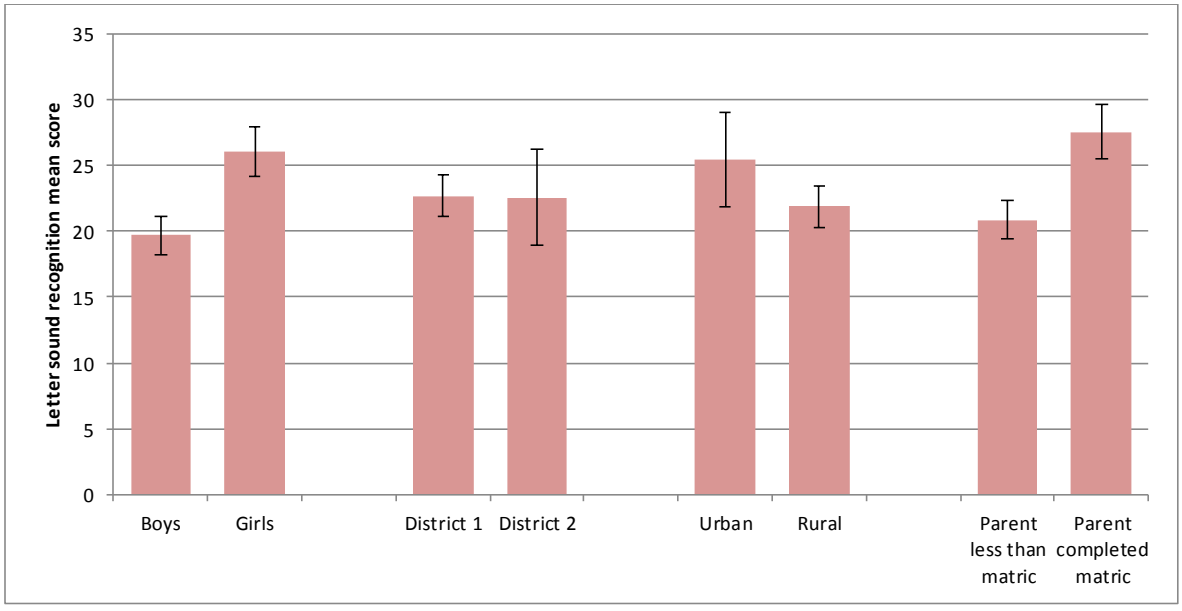
[9] Four of the phonological awareness items from the Baseline assessment were also administered in the Midline assessment.

Table 8: Correlation matrix of Midline sub-tests, Baseline and Wave 3 composite scores

	[A]	[B]	[C]	[D]	[E1]	[E2]	[F]	[G]	[Y1]	[Y2]	[Y3]
[A] Letter recognition	1										
[B] Word recognition	0.71	1									
[C] Non-word recognition	0.68	0.91	1								
[D] Sentence reading	0.65	0.72	0.73	1							
[E1] Paragraph reading	0.65	0.89	0.89	0.73	1						
[E2] Paragraph comprehension	0.64	0.85	0.86	0.80	0.88	1					
[F] Writing	0.62	0.60	0.57	0.65	0.59	0.61	1				
[G] Phonological awareness	0.55	0.62	0.61	0.59	0.62	0.64	0.55	1			
[Y2] Wave 2 composite	0.80	0.93	0.92	0.86	0.92	0.92	0.75	0.75	1		
[Y1] Baseline composite	0.20	0.24	0.23	0.23	0.22	0.23	0.19	0.17	0.25	1	
[Y3] Wave 3 composite	0.68	0.63	0.60	0.66	0.61	0.62	0.64	0.52	0.72	0.22	1

Figure 7 shows mean letter sound recognition for the same four subgroups as was presented earlier for baseline scores. The gender gap is now more pronounced than it was at the start of grade 1 with girls now being able to read about 6 more letters in a minute than boys, on average. A similar size gap exists between children who live with a parent or guardian with at least complete secondary school and children who do not. The two districts in the sample achieved roughly the same level of performance at the end of grade 1. Interestingly, there was no significant difference on average between the schools located in urban township areas and those located in deep rural settings.

Figure 7: Mean scores by sub-groups at Midline



Note: 95% Confidence intervals are shown

The midline data enabled us to estimate the impacts of each intervention after a single year of implementation during the course of grade 1. Strictly speaking, implementation of interventions only began at the start of the second school term of 2015, since the first term was taken up with baseline assessments, training teachers and other preparatory activities.

Table 9 reports the baseline and midline mean scores by treatment group, for the combined score as well as for the letter recognition sub-tests. In order to make this more visually accessible we also present the letter recognition scores using a percentile plot. Figure 8 shows this only for the control group and Intervention 2 since this is the group that exhibited the largest gains. The gains relative to the control group seems to be the largest at around the median. Whereas both treatment 2 and control groups median baseline number of letters correct was about 2 letters, by midline the median treatment 2 learner could recognize 23 letters in a minute compared to 16 letters in the control group.

Table 9: Simple comparison of baseline and midline mean scores

	(1) Baseline combined score	(2) Endline combined score	(3) Baseline letter recognition	(4) Endline letter recognition
Control	0.0404	-0.0179	5.406	22.70
Treatment 1	-0.170	-0.00675	4.144	22.01
Treatment 2	0.108	0.0992	5.836	25.14
Treatment 3	-0.00172	-0.0644	4.740	20.79
Observations	4538	4143	4538	4143

Figure 8: Percentile plot of letter recognition at baseline and midline for Treatment 2 and Control

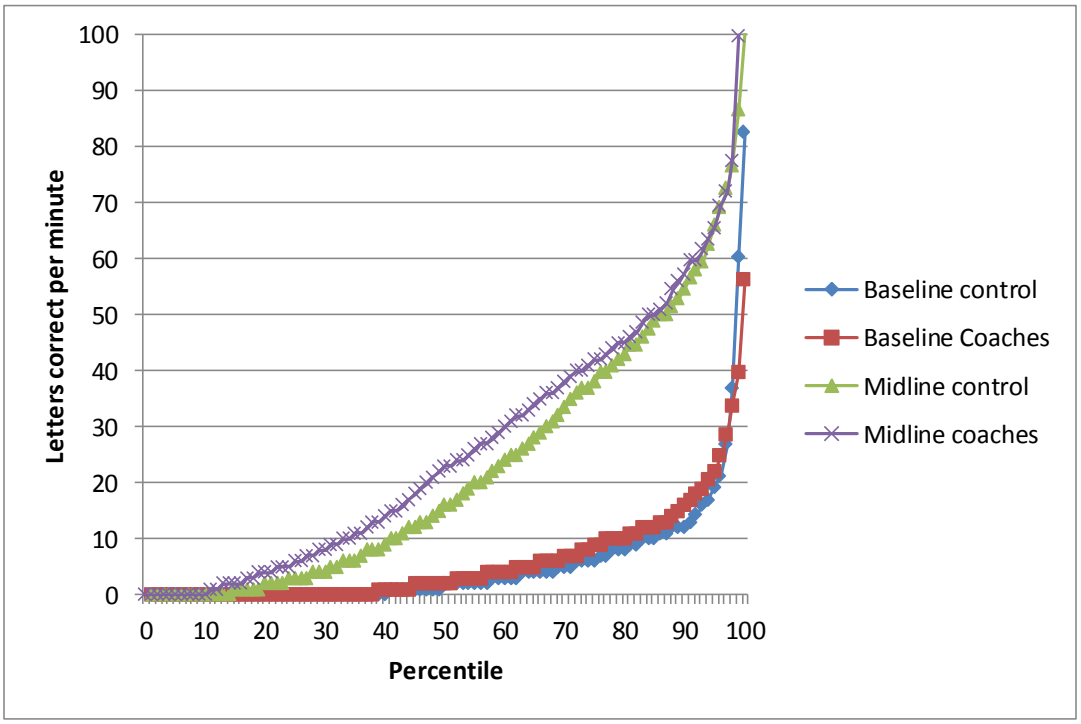


Table 10 presents the results from our main regression model specification applied to the midline impact evaluation. This model controls for baseline scores, district (schools are spread randomly across two districts), school mean score in the Annual National Assessments of 2014 (the most recent standardized school assessment), learner gender, parent education (according to the parent/guardian questionnaire), and two community-level controls obtained from the national census of 2011, namely a community wealth index derived from several questions about household possessions and the proportion of 13 to 18 year-olds in the community that are attending an educational institution. The motivation for including these controls is to account for any incidental differences that may exist between the treatment groups as well as to improve the precision of the estimates by increasing the explanatory power of the model.

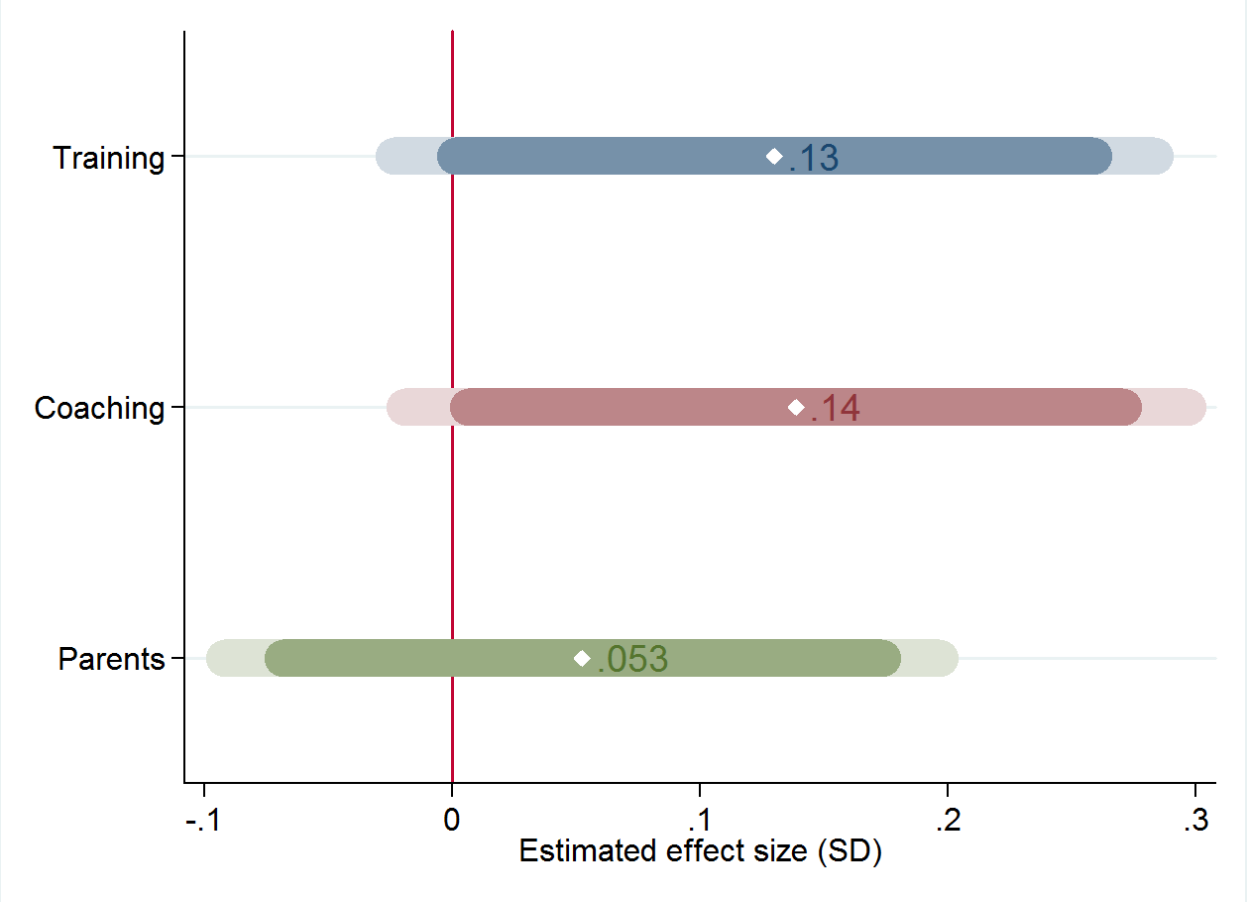
The coefficients on the dummies for Interventions 1 and 2 are both statistically significant, but only at the 90% level. Both of these coefficients are of a magnitude of about 0.13 to 0.14 standard deviations. The coefficient on the intervention 3 dummy (Parent Involvement programme) is neither statistically significant nor large enough to be considered educationally meaningful. The estimated treatment effects are graphically represented in Figure 9.

Table 10: Year 1 regression models with full controls (Main specification)

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.130*		
	(0.0777)		
Intervention 2		0.139*	
		(0.0799)	
Intervention 3			0.0526
			(0.0730)
Constant	-1.811***	-1.498**	-1.070**
	(0.560)	(0.578)	(0.475)
Observations	2,321	2,359	2,345
R-squared	0.190	0.208	0.243

Notes: Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 9: Graphical representation of estimated treatment effects showing 90% and95% confidence intervals



The midline results can be summarized as follows. We observed small to moderate positive impacts of both treatments 1 and 2 on reading outcomes at the end of grade 1. Using some rough calculations, assuming that the gains from baseline to midline in the control group are reflective of a year of learning, we estimate that the two pedagogical interventions yielded an impact of approximately 20% of a year of learning during three quarters of a year. However, the 90% confidence intervals ranged from slightly above zero to nearly twice that impact size.

Overall, the impact of the parent involvement intervention was small, and a zero impact could not be ruled out. The most likely reason for low average impact was low attendance rates amongst parents at the weekly meetings.

The impacts of treatments 1 and 2 were clearer for boys than for girls. For boys, each of these interventions had an estimated effect of 0.19 standard deviations, and this was statistically significant at the 95% level of confidence. This result could be a positive finding for the sake of helping boys catch up to girls in literacy outcomes.

The positive estimated effects of treatments 1 and 2 were clearer amongst schools in urban areas (33% of our sample), where the estimated effects were higher than 30% of a standard deviation. Consistent with this, amongst schools classified as Poverty Quintile 1 (which are poor and rural) there appeared to be no impacts of treatments 1 and 2. Finally, when excluding the few multi-grade schools in our sample, the positive impacts of treatment were also clearer. There was some evidence that treatment effects were larger in schools where teachers were less frequently absent from school. We also observed that children in the upper middle range of achievement stood to benefit most from the pedagogical interventions, possibly indicative of the level at which the lesson plans were pitched. However, no children experienced negative impacts.

We also observed some evidence of shifts in intermediate outcomes in the form of changed teacher and classroom practice. We found that grade 1 teachers in interventions 1 and 2 were more likely to “stream” children into groups according to their reading proficiency, compared to the control group. Treatment 2 teachers appeared to conduct individualized reading assessments of learners more frequently than the control group. There was some evidence of increased reading resources in treatment 1 and 2 classrooms, especially of Setswana posters. Encouragingly, based on an inspection of learner exercise books, there was consistent evidence of more exercises of all types (including drawing pictures), of written exercises, and of full sentence writing exercises in both treatment 1 and 2 schools compared to the control group.

10. MONITORING THE FIDELITY OF INTERVENTION IMPLEMENTATION (YEARS 1 & 2)

INTERVENTION ONE

Although 50 schools were randomly selected for this programme, one school fell out of the programme because it was merged with another school and thus no longer existed, leaving 49 schools actively participating. All 50 schools were, however, tested as part of the evaluation side of the project. In 2015, grade 1 teachers received the intervention, while in 2016 grade 2 teachers in the same schools were targeted. Teachers were provided with teaching and learning materials and were trained on their use twice each year for two days at a time in a conventional model common to large-scale training. The teachers received the official support of the North West Department of Education, and Class Act was in contact with them during the year through social media platforms.

The table below details the teaching and learning materials that teachers were provided with during the project.

Table 11: Intervention 1 materials

ITEM	DESCRIPTION
Vula Bula Reading Books	Commercially produced Grade One and Grade Two Setswana graded reading books. These were used in group guided reading lessons
Book register	An exercise book set up as an accession register for the Vula Bula reading books
Teacher file	A management file to keep teaching and learning materials
Setswana HL scripted lesson plans	This document contains the individual lesson plans that teachers followed in 2015 and in 2016
Flashcard words	Printed sets of the words teachers needed to teach sight words in reading lessons
Reading words	A learner resource that listed the sight words taken from the Vula Bula books. These word lists were taken home so that the learners could practice reading
Assessment records	CAPS and SA-SAMS compliant assessment record tables. Teachers used this resource to record formal assessments per learner
Assessment rubrics	Criteria for teachers to use to award objective assessment ratings for learner tasks
Curriculum tracker	A tool for teachers to manage curriculum coverage
Weekly routine	A tool for teachers to manage curriculum pacing
Core methodologies	Detailed pedagogical support that helped teachers learn how to use tried and tested methodologies for different language components
Handwriting poster	A poster that demonstrated the form and directionality of lower and upper case letters
Theme posters	Posters that detailed interesting scenes that were used for vocabulary development
Facilitators' Guides	Detailed handbooks for trainers to follow when they trained teachers.

The 2-day training events occurred four times over the course of 2015 and 2016 and were well attended, as can be seen in Table 12. The first training event in February 2015 covered the lesson plans for Term 2 only (and not Term 1) since the learning programme only began in Term 2 due to the other preparatory activities taking place in Term 1. This would not have

meant any disruption in the learning programme in intervention schools since the lesson plans are aligned to the official curriculum, only specifying a greater level of detail and with particular activities, instructional methods and resources integrated into the lessons. Only on one occasion was a school not represented at all at a training session. The attendance rate for teachers was also high (between 85% and 100%) and was sustained throughout the two years. Teachers who did not attend the residential training sessions were provided with catch-up training. Attendance rates were a little lower for school leaders (principals or HODs) but this is not a major concern since they were not the primary recipients of training.

Table 12: Attendance rates at training events

	GRADE ONE (2015)		GRADE TWO (2016)	
	ACTUAL TERM 2 FEB 2015	ACTUAL TERM 3 & 4 JULY 2015	ACTUAL TERM 1 & 2 JAN 2016	ACTUAL TERM 3 & 4 JULY 2016
% Schools attended	100	98	100	100
% Teachers attended	100	85	98	93
% School leaders attended	74	78	90	80

In 2016 a series of visits to schools and phone calls were conducted in order to encourage schools to remain committed to the programme. The use of phone calls and social media communication with teachers are cost-effective strategies that could relatively easily be added to conventional teacher training activities to encourage implementation. The use of visits is perhaps more costly to conduct at a large scale. In this regard, Intervention 1 might be thought of as a best case scenario of the conventional teacher training model. Table 13 reports the percentage of schools that received visits and phone calls during 2016.

Table 13: Motivational visits / phone calls to schools

	GRADE ONE (2015)				GRADE TWO (2016)			
					VISITS		PHONE CALLS	
	TERM 1	TERM 2	TERM 3	TERM 4	TERM 1	TERM 2	TERM 3	TERM 4
% VISITS	N/A	N/A	N/A	N/A	94	63	49	49

The service provider for this programme, Class Act, offered the following reflections regarding the effectiveness of Intervention 1:

TRAINING DOSAGE

- Residential training was held twice in each year (Term 1 and Term 3). This was very effective as trainers had significant amounts of time to work with teachers.

- However, the six month period between contact sessions was problematic.

- In future initiatives it is suggested that this form of training is supplemented with shorter one day contacts at the start of Term 2 and Term 4, although this would add somewhat to the cost of the intervention model.

TRAINING PARTICIPANTS

- In addition to teachers, the residential training included subject advisors and principals as participants.

- This was effective in that subject advisors were well trained, but it did open up the possibility of contamination to other schools due to official take-up of the project methodologies and materials.

- In some instances including principals was less effective as they did not engage with the actual training and used the time to discuss other issues with their peers. This was predominantly the male principals who apparently do not give high status to the Foundation Phase in their schools.

TRAINING CONTENT

- The first training session each year dealt with content and methodological issues directly related to Home Language (Setswana).

- In the second training session, these issues were deepened and then time was spent on lesson demonstrations by participants. This was very effective as teachers discussed and reflected on their lessons under the guidance of the trainers.

- In addition, teachers brought examples of learner work to the second training sessions. This work was presented as a gallery walk and proved to be a powerful motivational mechanism.

INTERVENTION TWO

Intervention Two also targeted teachers (Grade 1 in 2015; Grade 2 in 2016), but in a different randomly selected group of schools across the same districts as Intervention One. One school fell out of the programme because it was a multi-grade school and the principal therefore requested not to be part of the project since we were using grade-specific lesson plans. This left 49 schools actively having participated. All 50 randomly selected schools were still included in the data collection for evaluation purposes. These teachers were provided with the same teaching and learning materials as Intervention One (see Table 11 in the previous section). But

they received more intensive cluster-based training four times a year and had the support of instructional coaches in their schools and in their classrooms. The teachers also received the official support of the North West Department of Education and were in ongoing contact with their coaches between support visits throughout the year through social media platforms.

Table 14 shows how the coaches were allocated to schools. The schools were divided geographically across the three coaches. Many schools had more than one teacher to support. During 2016 one teacher retired which reduced the original number of Grade 2 teachers to 88.

Table 14: Allocation of reading coaches

COACH	DISTRICT	GRADE ONE (2015)		GRADE TWO (2016)	
		NUMBER OF SCHOOLS	GRADE 1 TEACHERS	NUMBER OF SCHOOLS	GRADE 2 TEACHERS
Coach 1	Ngaka Modiri Molema	17	27	17	26
Coach 2	Ngaka Modiri Molema	18	34	18	32
Coach 3	Dr Kenneth Kaunda	14	34	14	31
TOTAL		**49**	**95**	**49**	**89**

Table 15 summarizes the attendance of teachers at the various training engagements as well as the dosage of on-site coaching visits. High attendance levels were noted throughout the project demonstrating ongoing commitment. Teachers were supported throughout the project in their classrooms between 2 and 3 times per term. Fewer coaching visits were possible per teacher in the last term of 2016 due to a combination of social unrest in one district and learner assessments and other outside disruptions in both districts. In addition to classroom-based support, teachers received additional support during needs-driven afternoon workshops amongst nearby clusters of schools, which were facilitated by coaches. Although these workshops did happen to some extent in Year 1, these support initiatives became more structured in Year 2 and were reported on in Year 2. Due to the shortened length of Term Four and due to disruptions to schooling in the area no afternoon workshops were run by coaches.

Table 15: Summary of attendance and dosage of Intervention Two

	GRADE ONE (2015)			GRADE TWO (2016)			
	TERM 2 FEB 2015	TERM 3 JUL 2015	TERM 4 SEP 2015	TERM 1 JAN 2016	TERM 2 APR 2016	TERM 3 JUL 2016	TERM 4 SEP 2016
% Schools attended 1-day training	100	92	100	100	100	96	100
% Teachers attended 1-day training	100	89	100	99	100	92	99
Average number of on-site coaching visits	3	2	2	2	3	3	1
% Teachers attended cluster-based afternoon workshops				48	59	61	0

Two further indicators of programme implementation by teachers were monitored in Intervention Two schools: curriculum coverage and learner assessment records. These could not be monitored in other schools since the service provider only made routine visits to classrooms for Intervention 2 by the nature of the intervention. Coaches used the curriculum trackers and the work in learners' exercise books to ascertain the level of curriculum coverage across each language component area. The high levels of curriculum coverage reported in Table 16 are high relative to what other studies have observed, and relative to Class Act's experiences in other projects.

Table 16: Curriculum coverage in Intervention 2 schools

	GRADE ONE (2015)					
	PHONICS	HAND WRITING	LISTENING & SPEAKING	SHARED READING	GUIDED READING	CREATIVE WRITING
% COVERAGE	82	82	77	77	66	74
	GRADE TWO (2016)					
	PHONICS	HAND WRITING	LISTENING & SPEAKING	SHARED READING	GUIDED READING	CREATIVE WRITING
% COVERAGE	83	81	86	85	75	85

The learner assessment requirements reported on in Table 17 were taken from the formal assessment tasks required by CAPS. The results documented come from the teachers' assessment records that were based on criterion-referenced instruments supplied with the EGRS project materials.

Table 17: Assessment records from Intervention 2 schools

	GRADE ONE (2015)					
	PHONICS	HAND WRITING	LISTENING & SPEAKING	READING	CREATIVE WRITING	AVERAGE
% RESULTS	63	66	64	63	66	64
	GRADE TWO (2016)					
	PHONICS	HAND WRITING	LISTENING & SPEAKING	READING	CREATIVE WRITING	AVERAGE
% RESULTS	71	64	65	70	59	64

Class Act offered the following reflections regarding the effectiveness of Intervention 2:

TEACHER TRAINING AND DEVELOPMENT

- Cluster-based training was held four times each year to prepare teachers for the term ahead. This was 1-day training held at schools and therefore did not involve any accommodation costs.

- This was supplemented by needs-driven training that was offered during the terms by the instructional coaches.

COACHING

- Trusting and meaningful relationships were built between coaches, teachers, principals and school administration staff.

- The coaches managed to balance these relationships with appropriate measures of familiarity and professionalism.

- As such, accountability measures were used in conjunction with praise, motivation and positive reinforcement to enhance teacher development and learner outcomes.

DIFFERENTIATED COACHING

- For the first three terms of each year all teachers received the same amount of individual coaching sessions (between 2 and 3 per term).

- However, for Term 4 in 2015 and 2016, teachers who were deemed to be coping well received less coaching than teachers who were deemed to be trying but struggling.

- There were some teachers who, due to their lack of commitment to the programme, received reduced levels of coaching as well.

- The process of instructional coaching assists in analyzing the extent to which individual teachers need support, and the extent to which individual teachers will take-up the support offered.

CONTEXTUAL FACTORS

- The broad context of the school and the teachers impacted on the extent to which instructional coaching could prove to be effective. For example:

    - Social unrest in Ngaka Modiri Molema impeded progress in Term 4 2016 as teachers could only begin teaching in Week 4.

    - Many schools were difficult to access in the rainy seasons in both years, resulting in coaches making alternative arrangements for off-site or telephonic support.

- In some small schools the Principal was either the Grade One or Grade Two teacher resulting in some coaching visits being cancelled and / or postponed due to management responsibilities.

- Teacher turnover was high in four schools in particular. This turnover ranged from one school having no teacher at all until DBE intervention to other schools having different teachers each term.

- Absenteeism was noted as a coping mechanism by some teachers who wanted to avoid the accountability associated with a visit from a coach.

TEACHING AND LEARNING MATERIALS

- The lesson plans and additional resources provided by the project were effective tools in the acquisition of reading skills.

- In both years teachers grasped the technical methodologies early and eagerly (especially phonics and handwriting).

- It was noted that many teachers initially confused shared reading and group guided reading methodologies.

- Process writing proved to be a difficult methodology for teachers more comfortable with a product view of creative writing.

INTERVENTION THREE

Intervention Three was conducted in a third randomly selected group of 50 schools. One school fell out of the programme as it was a boarding school and the parents stayed too far away to attend weekly meetings. Therefore, 49 schools actively participated in the programme, although some of these schools experienced disruptions to the programme due to delays in finding a Community Reading Coach (CRC) to run the weekly parent meetings. This intervention focused on parents/guardians (Grade 1 in 2015 and Grade 2 in 2016) through weekly meetings facilitated by CRCs. The weekly sessions were aimed at helping parents understand how their children were learning to read and to provide parents with strategies to use at home to stimulate a culture of reading. CRCs were recruited with the help of school principals and received regular training sessions from Class Act (6 sessions in 2015 and 7 in 2016). The CRCs were in contact with Class Act during the project through social media platforms. The table below details the materials that parents were provided with during the project. For each module, three meetings were held, each covering the same topic but using a different set of activities. This configuration meant that parents could attend once every three weeks and still have “full

coverage", but that if parents attended every week they would not repeat the exact same session.

Table 18: Intervention 3 materials

ITEM	DESCRIPTION
Module One	'Small things can make a difference' plus a set of family reading cards
Module Two	'Playing with sounds to support reading' plus a set of family reading cards
Module Three	'Reading pictures' plus a set of family reading cards
Module Four	'Letter sounds' plus a set of family reading cards
Module Five	'Incidental reading' plus a set of family reading cards
Module Six	'Preparing to read a story – Part One' plus a set of family reading cards
Module Seven	'Reading a story' plus a set of family reading cards
Module Eight	'Preparing to read a story – Part Two' plus a set of family reading cards
Module Nine	'Reading Remediation'
Module Ten	'Reading stories 3 and 4'
Module Eleven	'Reading story 5'
Module Twelve	'Reading stories 6, 7 and 8'
Module Thirteen	'Reading stories 9, 10 and 11'
Module Fourteen	'Reading stories 12, 13 and 14'
Module Fifteen	Consolidation
Grade One reader	Platinum Series Le Re Tlhabetse Readers published by Maskew Miller Longman / Pearson: Book 1
Grade Two reader	Platinum Series Le Re Tlhabetse Readers published by Maskew Miller Longman / Pearson: Book 2
Facilitators' Guides	Detailed handbooks for trainers to follow when they trained CRCs. These were also used for parent training.

CRC attendance at training sessions was mostly fairly high, as Table 19 shows, and where CRCs did not attend catch-up training was provided as far as possible.

Table 19: CRC attendance at training

	APR 2015 MOD 1	MAY 2015 MOD 2&3	JULY 2015 MOD 4	SEPT 2015 MOD 5&6	OCT 2015 MOD 7	NOV 2015 MOD 8	
%	100	90	71	86	90	82	
ATTENDANCE	JAN 2016 MOD 8 REVISION	MAR 2016 MOD 9&10	APRIL 2016 MOD 11	MAY / JUNE MOD 12	JULY 2016 MOD 13	SEP 2016 MOD 14	OCT 2016 MOD 15
	98	90	90	86	73	75	78

CRCs kept attendance registers of parents who attended the weekly meetings. Parents signed against the learner's name so as to be able to track attendance relative to the learner. This was done with a view to linking the attendance data to the learner test data collected independently. Table 20 shows the percentage of learners who were represented at the meetings. The percentage who attended at least one session per module is shown. The average parental involvement throughout the duration of the project was approximately 31%. This low level of

parental involvement remained a concern across both years despite attempts to address this during Year 2.

Table 20: Percentage of learners whose parent/guardian attended meetings

	MOD 1	MOD 2&3	MOD 4	MOD 5&6	MOD 7	MOD 8	
% ATTENDANCE	42	36	26	38	31	22	
	MOD 9	MOD 10	MOD 11	MOD 12	MOD 13	MOD 14	MOD 15
	36	35	34	26	25	19	37

Class Act supervisors conducted monitoring visits of parent meetings at 33 of the 50 schools and evaluated the effectiveness of the meetings. Table 21 shows the numbers of monitoring visits falling into the various categories of assessed effectiveness. The reasons for classifying a school as "not working" included a combination of the following:

- The principal was unsupportive.
- The CRC did not adhere to scheduled parent meetings.
- The school did not appoint a CRC.
- Very few parents attended the CRC session.

Table 21: Monitoring visits to weekly parent meetings conducted in 2016

	ACTUAL VISITS	WORKING WELL	WORKING SATISFACTORILY	NOT WORKING
Number of visits	33	5	17	11
%	67	15	52	33

Class Act offered the following reflections regarding the effectiveness of Intervention 3:

COMMUNITY READING COACHES

- During the project more regular contact and training sessions were implemented between the master trainers and the Community Reading Coaches (CRCs).

- Repetition of key issues and developmental opportunities effectively prepared the CRCs for parent contact sessions.

PARENTAL INVOLVEMENT

- Low levels of parental involvement (average of 31%) remained a challenge throughout the project.

OVERALL DESIGN

- Principals and Subject Advisors have authority over teachers, thus teacher commitment and involvement is easier to secure and maintain (as compared to that of parents).

- Many parents still feel that educating their children is solely the responsibility of the school and of the teachers.

- Local politics interfered with the effectiveness of some CRCs where issues of patronage and protection were experienced.

11. YEAR 2 RESULTS

WAVE 3 INSTRUMENT DEVELOPMENT AND PILOTING

The Research Team worked closely with the HSRC to develop four survey instruments for the Wave 3 data collection: a learner test, a school principal questionnaire, a teacher questionnaire and a parent/guardian questionnaire. The learner test was designed in the spirit of the Early Grade Reading Assessment (EGRA) to be administered orally by a fieldworker to one child at a time. The test instrument used parts of the EGRA for Setswana, which had already been developed in South Africa. A Setswana linguist consultant (accredited assessor, teacher and translator) assisted the Research Team.

The Wave 3 instruments were piloted on 6-7 September 2016 in the same five schools where the piloting of the baseline and midline instruments had previously taken place. Further refinements to the instruments were then made by 26 September 2016 in preparation for the start of fieldwork during the final week of October 2016. The training and administration manual was also revised to ensure alignment to the Wave 3 instruments. Eight sub-tests were included in the final Wave 3 learner assessments. They were:

- EGRA Letter Sound Recognition
- EGRA Word Recognition
- EGRA Non-Word Decoding
- Paragraph Reading (Oral Reading Fluency)
- Reading Comprehension
- Writing
- Phonemic Awareness.
- Mathematics
- English vocabulary

The piloting indicated that each individual oral assessment should be completed within 15 minutes per learner.

ETHICAL CLEARANCE

The HSRC's Research Ethics Committee (REC) approved the initial project design on 24 March 2014. The baseline instruments along with an application for recertification for another year were approved on 21 January 2015. The pilot-versions of the Wave 2 instruments, information sheets and consent forms, as well as procedures, were approved on 2 September 2015. The final Wave 2 data-collection versions of the foregoing, including approval of a deviation request to involve learners and staff from five additional schools in simulation training, were granted clearance on 13/14 October 2015. Recertification of the study for 2016 was approved on 6 November 2015.

The intended instruments related to Wave 3 data collection were formally submitted to the HSRC's Research Ethics Committee on two occasions: before piloting them in September 2016, and before conducting the training and administering the instruments in October 2016. Clearance for the Wave 3 piloting administration was provided on 2 September 2016, and for the Wave 3 main (quantitative) data collection from 24 October to 11 November 2016, on 11 October 2016. The latter included registering and obtaining clearance for the deviation request to allow for the following two parts of the additional qualitative research:

(a) Classroom observation instrument set (teacher interview, classroom observation, and document review) to be administered at a sub-sample of 60 schools from 13-26 October 2016; and

(b) "In-depth" qualitative lesson observation instrument to be administered at a sub-sample of 4 schools from 17-20 October 2016.

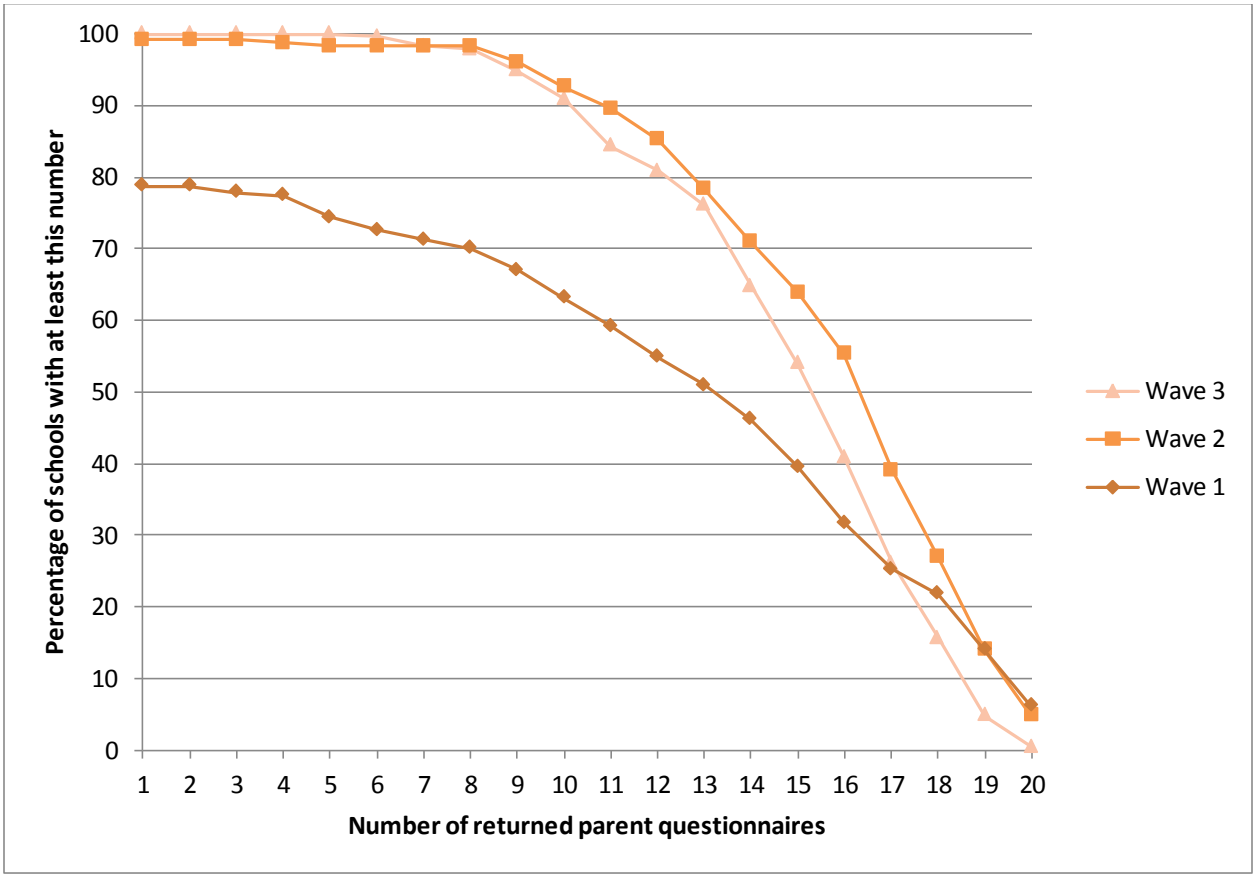
WAVE 3 DATA COLLECTION

The various measures that were put in place by the HSRC and the Research Team in the Midline survey (October/November 2015) were again applied in Wave 3 (October/November 2016) with similarly positive results in terms of instrument response rates and the general quality of work. One difference in the approach to Wave 3 was that a deviation from the usual HSRC procurement procedures was granted in order to automatically re-hire the same data collection organisation that had undertaken the Midline survey. This saved time, effort and money, and was done because of the high quality of work that had been provided at Midline.

The response rates to the parent, teacher and principal questionnaires at baseline were disappointing. The parent questionnaire was sent home with tested children and was meant to be brought back to the school and then collected on a later day by the fieldwork agency. The weakness of this method is that children may not always bring the questionnaire back. However,

it is more reliable than asking young children themselves about home characteristics, such as their parents' education. At baseline there were 49 schools for which no parent questionnaires were returned. At midline this was the case for only 2 schools while in Wave 3 all schools returned at least 5 completed parent questionnaires. At baseline roughly 40% of schools had fewer than 10 completed parent questionnaires. In contrast, in Waves 2 and 3 only about 10% of schools had fewer than 10 completed questionnaires. This is likely to reflect the difference in fieldwork quality between baseline and the subsequent data collections. Importantly, there no significant differences in instrument return rates across Intervention groups, not that it would be expected since fieldworkers were blind to Intervention allocation.

Figure 10: Percentage of schools with at least this number of parent questionnaires returned per school



Similarly, there were 32 schools at baseline for which no teacher questionnaires were returned. At midline, there were only 2 such schools. At baseline, we received teacher questionnaires for 326 teachers compared to 383 teachers at midline and 356 at Wave 3. At Wave 3, there were 7 schools for which no teacher questionnaires were completed. At baseline, there were 14 schools for which we received no principal questionnaires, compared to just 2 schools at midline and 3 at Wave 3.

DATA CAPTURING AND CLEANING

Once the completed survey instruments had arrived back at the HSRC, they were unbundled from their school batches and re-sorted by instrument or record type in preparation for data capturing and cleaning.

- Preparation of instruments for data processing

The data manager of the project, with two research trainees, went through the various instruments to ensure that correct coding and scoring had been accomplished. All queries were attended to before the instruments were batched and routed for data capturing. Given in-house capturing capacity constraints of the HSRC, permission and instruction was given for the data to be captured by an external service provider.

- Data capturing

The external data-capturing service provider was provided with all the record layouts, requirements and capturing templates for each instrument / dataset, and then trained for their task. On-going supervision and regular (virtually daily) checking of progress and quality were pursued by the HSRC's data manager.

The data-capturing service provider had to adhere to a process of 100% verification. This means that all data were captured twice; first into a temporary dataset and then, once the second capturing keystrokes were either identical, or a query had been solved on being discrepant, into the permanent output file.

Data capturing had been completed by January 2017, after which the preliminary datasets were provided to the Research Team in February 2017.

- Data cleaning and final hand-over

On receiving the initial datasets from the data-capturing service provider, the HSRC's data manager checked for double records, incorrect identity numbers, incorrect field values, and similar unexpected values and information, and consolidated such queried data-fields against the hard-copy completed instruments. Once these unexpected values and queries had been solved, and data labels and values completed, the dataset was provided to the Research Team on 14 February 2017, and after a few minor data queries were raised by the Research Team a final dataset was provided by HSRC on 15 March 2017.

ATTRITION AND GRADE REPETITION

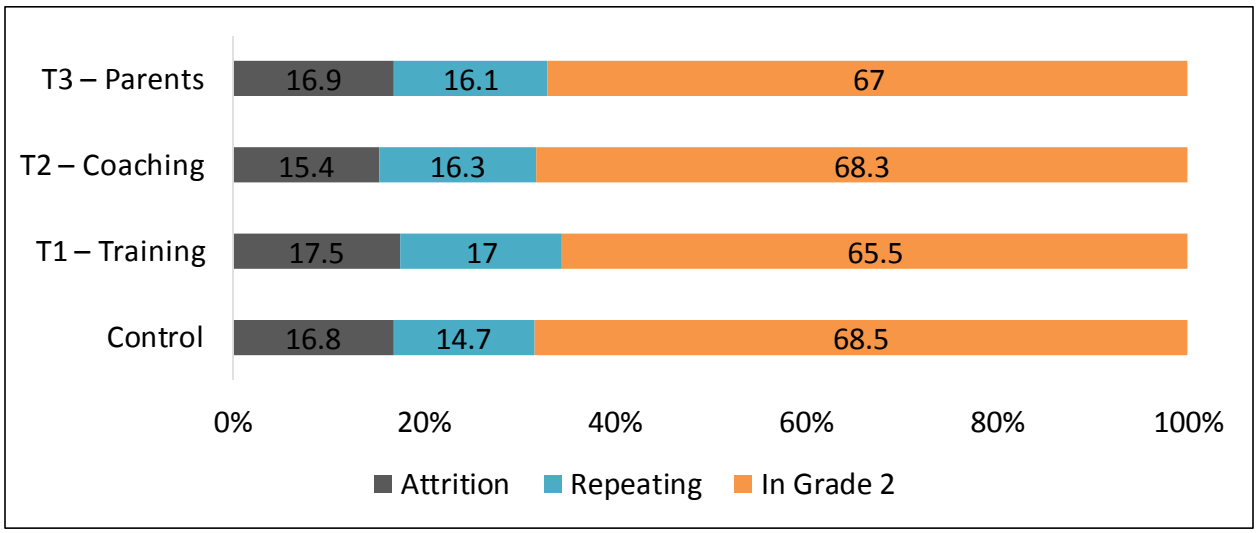
At baseline we assessed 4 538 learners at the start of their grade 1 year in 2015. At the midline assessment (end of grade 1) we successfully re-assessed 4 143 learners, meaning that about 9% of the original sample were either absent from school on the day of the survey, or had moved to another school, or had returned to Grade R, or had stopped attending school. However, at the end of grade 2, just over half of those of absent for the midline were successfully re-assessed, implying that they must have been absent from school on the day of the midline survey. Therefore, we can say that about 4% of the original sample appeared to have left the school before the end of grade 1.

By the time of the Wave 3 data collection (end of grade 2, 2016), we successfully assessed 3 781 learners (83.3% of the original sample). This means that, over and above the 4% of learners who left the school before the end of grade 1 in 2015, a further 13% were either absent on the day of the Wave 3 survey or had left the school since grade 1. Unfortunately, we are not able to distinguish between these two reasons.

Amongst those learners who were successfully identified at the Wave 3 survey, 591 were found to be repeating grade 1. This amounts to about 13% of the original sample, or about 16% of those successfully identified at Wave 3. The latter estimate is probably the more relevant estimate of the grade repetition rate for grade 1, though of course it is not based on a nationally or even provincially representative sample. Nevertheless, this estimate is in line with internal DBE analysis of LURITS data, and confirms that grade repetition is higher than what certain other sources, such as the General Household Survey, would suggest.

Figure 11 shows the proportions of the original sample that attrited, that were repeating grade 1, and that were found to be in grade 2. Importantly, there were no significant differences in attrition or grade repetition across the three intervention groups. Formal regression analysis confirmed this to be the case. This means that the tested samples of learners will not be selectively stronger or weaker in any intervention group, something that could have introduced bias into the impact analysis.

Figure 11: Attrition and grade repetition in the sample



It is interesting to investigate what factors *were* associated with attrition and with grade repetition. Table 22 shows linear probability regression models predicting whether a learner from the original sample was not found at the end of grade 2, either due to leaving the school or being absent for the survey. Model 1 controls for baseline achievement of learners, while model 2 controls for the achievement of the learner at the end of grade 1. In both models, prior achievement was not predictive of attrition. Therefore, we do not have a story of weaker learners dropping out of school before completing grade 2 (something we know is very rare in the early grades) or even of weaker learners being more likely to be absent. Several other factors were, however, associated with attrition. Boys were slightly more likely to attrit than girls. Older learners were more likely to attrit. Those in Quintile 1 schools (the poorest communities) were more likely to attrit than those in quintiles 2 and 3 schools. Interestingly, those in one district (Dr Kenneth Kaunda) were significantly more likely to attrit than those in the district of Ngaka Modiri Molema. It is difficult to know why this would be the case, but it was also found that grade repetition was higher in this district, after controlling for learner performance at the end of grade 1, which may point to this district having a stronger hand in recommending that children are held back from promotion. Some of those who either repeated grade R (in 2015) or were repeating grade 1 in 2016 may have been missed by the fieldworkers, thus contributing to attrition. Lastly, those in deep rural settings were more likely to attrit than those in urban areas.

Models 3 and 4 redefine attrition more strictly as those we assume left the school because they were neither found at Wave 2 nor Wave 3 (unless they happened to be absent from school at both Wave 2 and 3). Model 4 excludes those who were present at Wave 2 but not at Wave 3 since it is uncertain whether they were absent or had left the school, whereas Model 3 treats this group as “not attriters”. In models 3 and 4 there is no longer any relation between attrition and being in a rural school, implying that the higher attrition in rural schools evident in Models 1 and 2 was driven by absenteeism rather than by movement between schools. The other

predictors are broadly consistent across all 4 models indicating that movement between schools was more common among males, older learners and especially those in the district of Dr Kenneth Kaunda.

Table 22: Factors predicting attrition

VARIABLES	Model 1	Model 2	Model 3	Model 4
BL_totscore_SD	0.002		-0.002	-0.003
	(0.006)		(0.003)	(0.003)
female_W2	-0.021**	-0.011	-0.012**	-0.014**
	(0.010)	(0.010)	(0.005)	(0.006)
age_best_W1and2	0.027***	0.021**	0.010**	0.013**
	(0.009)	(0.009)	(0.005)	(0.006)
2.quintile	-0.039**	-0.023	-0.019***	-0.023***
	(0.015)	(0.014)	(0.007)	(0.008)
3.quintile	-0.036**	-0.032**	-0.008	-0.011
	(0.016)	(0.016)	(0.008)	(0.010)
1.district	0.086***	0.060***	0.037***	0.047***
	(0.021)	(0.020)	(0.012)	(0.014)
rural_dummy	0.048***	0.046***	0.009	0.014
	(0.017)	(0.016)	(0.010)	(0.012)
W2_LT_totscore_SD		-0.004		
		(0.006)		
Constant	-0.040	-0.031	-0.030	-0.045
	(0.062)	(0.062)	(0.030)	(0.036)
Observations	4,512	4,134	4,512	3,940
R-squared	0.018	0.011	0.022	0.026

Notes: Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 23 shows the factors predicting grade repetition based on a linear probability regression model. In this case, achievement on our reading assessment at the end of grade 1 turned out to be strongly predictive of repeating grade 1 the following year. This is a good sign as it indicates that the grade repetition decision is not a complete lottery, as some research has suggested it was in many schools (Lam *et al*, 2010). As one might expect given policy prescriptions, those who were already repeating grade 1 in 2015 did not again repeat grade 1 in 2016. Older learners were less likely to repeat, something which also makes sense given that maturity could be expected to be a relevant consideration in the decision to repeat, over and above learning progress. Neither the school's poverty quintile status nor the rurality of a school was predictive of grade repetition after accounting for the learning achievement of children. However, boys

were more likely to repeat even after controlling for performance on the reading assessments at the end of grade 1. We know from other data sources that boys in South Africa are more likely to repeat grades than girls, but this data shows that this pattern is not fully explained by the weaker learning achieved by boys. This is one advantage of having a longitudinal dataset. Similarly, grade repetition was higher in the district of Dr Kenneth Kaunda, even after accounting for learner performance.

Table 23: Factors predicting grade repetition

	Repetition
Wave 2 score	-0.099***
	(0.006)
Previously repeated Gr1	-0.037*
	(0.022)
Female	-0.031***
	(0.011)
Age	-0.053***
	(0.010)
Quintile 2	-0.008
	(0.020)
Quintile 3	-0.032
	(0.023)
District	0.047*
	(0.028)
Rural	-0.012
	(0.023)
Constant	0.523***
	(0.068)
Observations	3,511
R-squared	0.099

Notes: Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 12 below shows the grade repetition rates for boys and girls at various points in the reading achievement distribution. The graph indicates, for instance, that nearly 1 in 5 boys in the 4th decile of performance at the end of grade 1 repeated grade 1, whereas only about 1 in 10 girls at that same level of performance went on to repeat grade 1.

Figure 12: Grade repetition by performance and gender

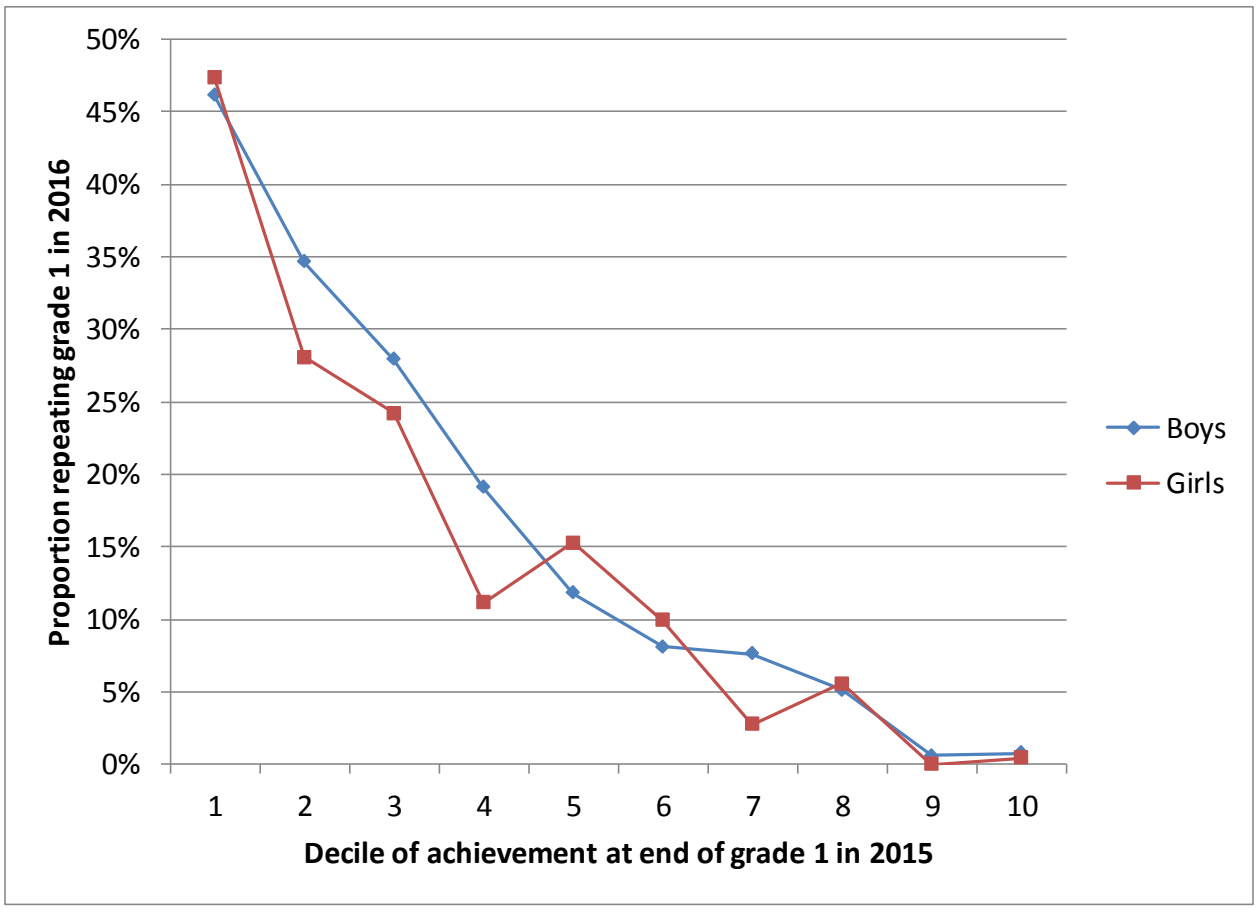
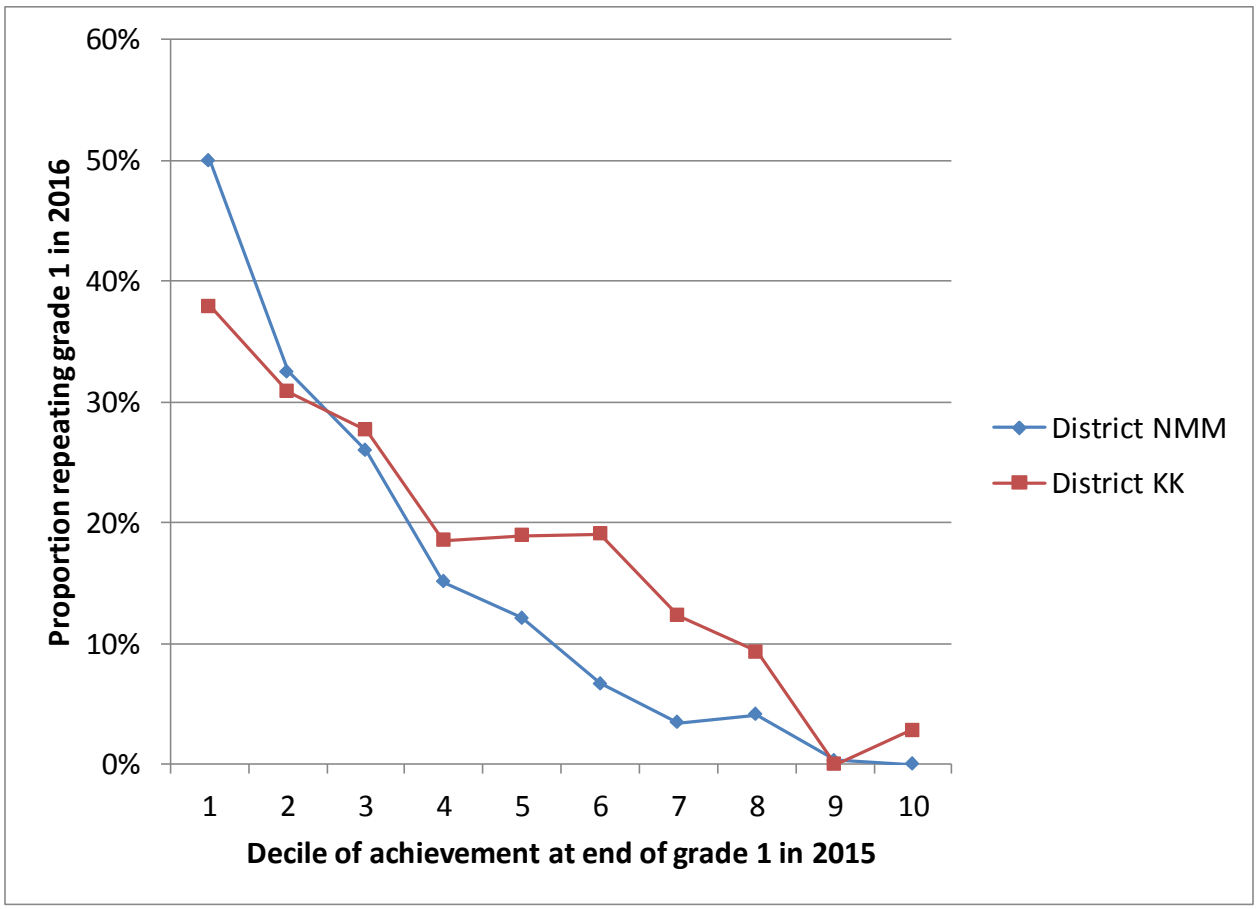


Figure 13 below shows the grade repetition rates for each of the two education districts at various points in the reading achievement distribution. The graph indicates, for instance, that children in Dr Kenneth Kaunda in the 6th decile of performance at the end of grade 1 were nearly three times more likely to repeat grade 1 than equally good readers in Ngaka Modiri Molema.

Figure 13: Grade repetition by performance and district



DESCRIPTIVE ANALYSIS OF LEARNER TEST RESULTS

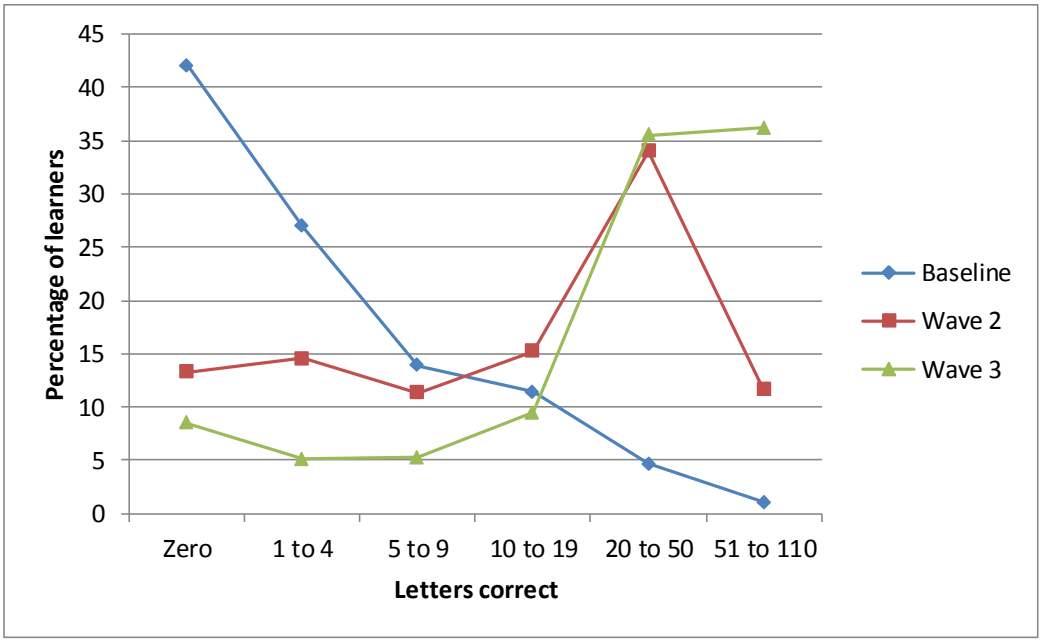
The learner assessment administered at the end of grade 2 (October/November 2016) was designed primarily to measure home language (Setswana) literacy outcomes, as this was the primary goal of each of our interventions. However, we included two grade-appropriate mathematics items and a few English reading items since these reflect the other two main learning areas within the Foundation Phase, namely mathematics and English as First Additional Language. Several teachers in the project had spoken about how the home language learning programme we were running was time-consuming and difficult to fit into the school timetable. Therefore, we suspected that it was possible that the EGRS home language literacy programme could have a negative effect on either mathematics or English through crowding out instructional time in those learning areas. Alternatively, we hypothesized that an effective home language programme could have positive spillover effects on mathematics or English through the acquisition of transferable skills or through the teacher's practice in one subject area improving due to training applied to another area. Therefore, we decided to include the mathematics and English items in the learner assessment.

The summary statistics for all the sub-tests are reported in Table 24. The first item, common to all three waves of data collection, was letter recognition. This item was only slightly adapted from the EGRA instrument developed in Setswana. Figure 14 shows learner performance on letter recognition at baseline, Wave 2 and Wave 3. At baseline, 42% of children had scored zero, as one might expect at the start of school. By the end of grade 1 there were about 13% of children who scored zero. Although there is clear improvement over time, it is concerning that there were still over 8% of learners who could not read a single letter after two years of schooling. If we exclude those repeating grade 1, then we can say that about 6% of the sample could not read a single letter at the end of grade 2. It is, however, encouraging that about 36% of learners could read at least 50 letters correctly in a minute by the end of grade 2.

Table 24: Summary statistics for each sub-test in Wave 3 learner assessment

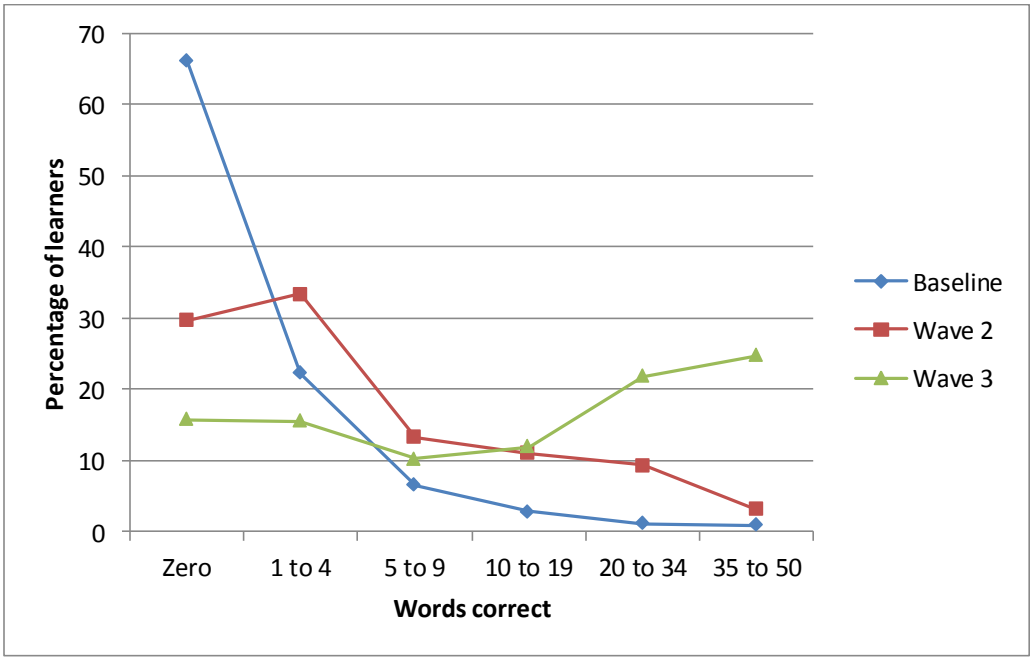
	count	mean	min	p10	p25	p50	p75	p90	max
Letter recognition	3781	39.5	0	2	16	41	60	74	110
Word recognition	3781	19.4	0	0	3	17	34	45	50
Non-word recognition	3781	14.4	0	0	0	13	26	34	50
Oral Reading Fluency	3781	25.6	0	0	0	23	50	64	66
Reading comprehension	3781	1.27	0	0	0	1	2	3	4
Phonological awareness	3781	1.82	0	0	1	2	3	3	3
Writing	3781	5.97	0	3	4	6	8	9	9
Mathematics	3781	0.6	0	0	0	1	1	1	2
English	3781	3.14	0	0	0	2	6	8	8
Composite score (SD)	3781	0	-1.59	-1.23	-0.97	-0.01	0.89	1.37	2.16

Figure 14: Letter recognition at Baseline, Wave 2 and Wave 3



The second sub-test, word recognition, was also a standard EGRA item and was common across all three waves of data collection. As one would expect, there was a big floor effect at baseline as well as at the end of grade 1. By the end of grade 2, about 16% of children could not read a single word, but there was also a substantial proportion (25%) who could read at least 335 words correctly in a minute.

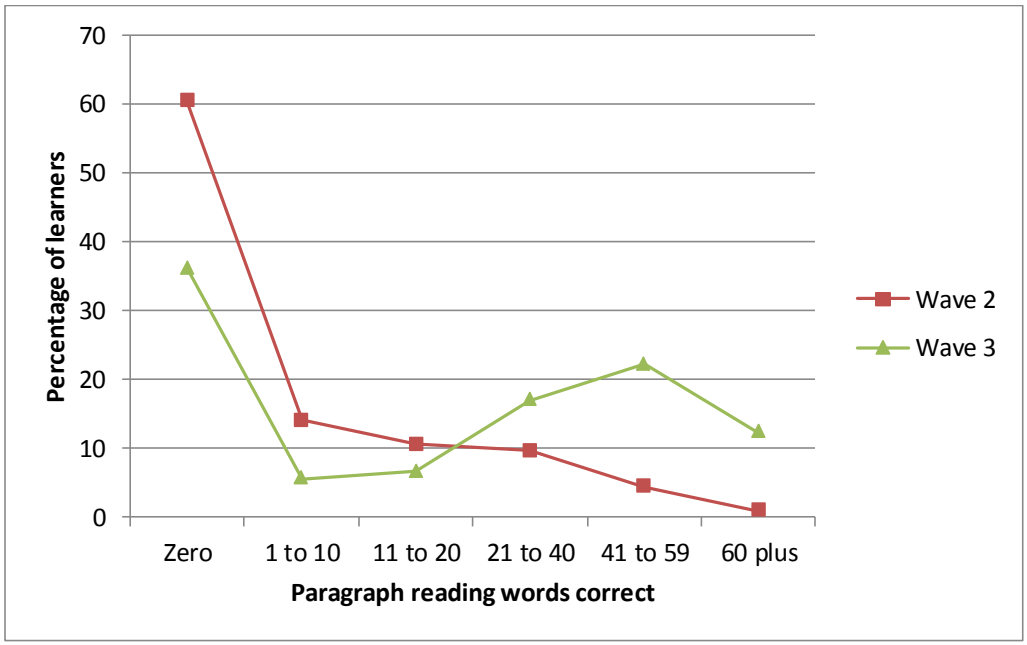
Figure 15: Word recognition at Baseline, Wave 2 and Wave 3



The third sub-test was non-word recognition. These are nonsense words which comprise letter combinations which are common to the Setswana language, but are not actual words. This tests decoding rather than whole word recognition. The scores tended to be somewhat lower than for word recognition.

The fourth sub-test was paragraph reading, which is really a measure of Oral Reading Fluency. Interestingly, for this item there were many learners who could not read a single word (36%) as well as many learners who scored highly, but few learners who read between 1 and 20 words correctly in a minute, as can be seen in Figure 16. This may suggest that once learners master the skill of decoding, reading fluency quickly follows. After reading the paragraph, learners were asked four comprehension questions about the passage. As one would expect, there was a strong correlation between oral reading fluency and comprehension scores.

Figure 16: Paragraph reading (ORF) in Wave 2 and Wave 3



Three items were included to tests phonological awareness. Two of these items asked learners to identify where in a spoken word a particular sound was located – at the beginning, in the middle or at the end. The third items asked learners to join together two sounds spoken by the fieldworker to form a single word. Although phonological awareness is a critical ingredient in the development of reading, one challenge associated with testing this skill is that fieldworkers (and hence learners) sometimes struggle to understand what is being asked of them, or to understand the difference between syllables and distinct sounds.

Several writing items were included in the Wave 3 test. One advantage of this sub-test is that there was no floor effect, with the score at the 10th percentile being 3 out of a possible 9 marks. First, learners had to write a single letter that was dictated to them, followed by a single word

dictated to them. Then, learners had to fill in a missing word in a short sentence corresponding to a picture. Finally, learners were asked to rewrite a 3-word sentence provided to them but with the correct punctuation. In this last item marks were awarded for each word, for placing spaces between the words, for using a capital letter at the start of the sentence and for placing a question mark at the end of the sentence.

As discussed earlier, two mathematics items (taken from a previous Annual National Assessment paper for grade 2) were included in order to assess whether the interventions had any positive or negative spillover effects. The first item was "8 + 3 = ___", and the second item was "halofo ya 28" which means "half of 28". The average score was 0.6 out of 2.

Similarly, a few English items were included. The first four items were single English words for the learner to read. Thereafter, the learner was asked to read a short English sentence consisting of four words. The average score was 3.14 out of 8.

We also constructed a composite score from all the home language literacy items using Principal Components Analysis (PCA). This is a type of factor analysis, which identifies the common underlying variation amongst a set of variables and regards the first principal component in this variation as being reflective of the underlying construct, which in our case would be Setswana reading literacy. For ease of interpretation we then standardized the composite index to have a mean value of 0 and a standard deviation of 1. In much of the literature reporting on Early Grade Reading Assessments it is not common practice to derive a composite score, but rather to interpret changes in each of the sub-tests separately. To some extent this is motivated by a theoretically driven approach to analyzing the development of reading and to some extent by the idea that the various sub-tests do not easily fit into a single underlying construct. However, we were also concerned about the risks of cherry-picking results if we present impact analysis on each of seven outcomes and then for all of these outcomes also go on to present heterogeneous treatment effects. Therefore, we have opted to derive the composite score and use this in all our heterogeneous treatment effect estimations and robustness checks. However, for the main impact evaluation model we also present the results on each of the sub-tasks separately.

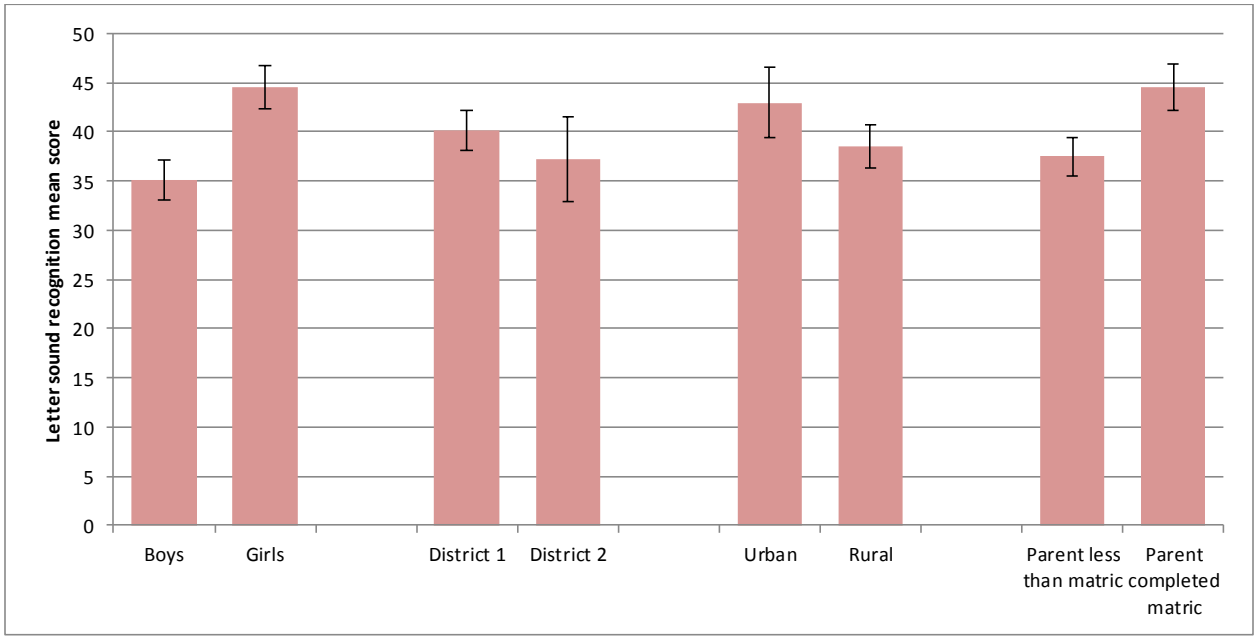
Overall, the learner test information from Waves 2 and 3 appears to have been of a better quality than that obtained through the baseline. This is evident in the Cronbach's alpha values which were 0.83 for both the Wave 2 and Wave 3 learner tests, but only 0.65 for the baseline assessment. Cronbach's alpha provides a measure of how well the various sub-tests fit together as measures of a single underlying construct, where a value closer to 1 is better. Table 25 below also provides evidence of how the baseline test provided a weak signal of learner proficiency. While the Wave 2 and Wave 3 total scores are highly correlated (0.72) each of them is only weakly correlated with the baseline total score.

Table 25: Correlation coefficients between Wave 1, 2 and 3 scores

	Baseline Total Score	Wave 2 Total Score	Wave 3 Total Score
Baseline Total Score	1		
Wave 2 Total Score	0.25	1	
Wave 3 Total Score	0.22	0.72	1

Figure 17 shows mean letter recognition scores at Wave 3 for several sub-groups of interest – letter recognition for the same sub-groups was presented for the baseline and Wave 2 scores earlier. Here it is evident that overall achievement is not significantly different between the two districts and nor is it significantly different between urban and rural schools. However, the gap between girls and boys has kept growing since Wave 2 with girls now reading nearly 10 more letters correctly per minute than boys. Similarly, children with parents who have at least matric (complete secondary school) could read about 8 more letters correctly per minute than children whose parents were not as educated.

Figure 17: Mean scores by sub-groups at Wave 3



DESCRIPTIVE ANALYSIS OF TREATMENT EFFECTS

A simple comparison of means provides an initial perspective on the possible impact of the three interventions, as reported in Table 26. Glancing over the table indicates that children in the Intervention 2 group had consistently higher average scores than the other groups. The scores for Interventions 1 and 3 do not appear noticeably higher than the control group

(although one should keep in mind that Intervention 1 had lower baseline average achievement, as the balance tests revealed).

Table 26: Mean scores for all sub-tests by intervention group

	Control	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Letter recognition	39.04	37.52	43.01	38.70
Word recognition	18.91	18.77	22.28	18.02
Non-word recognition	13.69	13.87	16.99	13.40
Oral Reading Fluency	24.48	24.95	29.90	23.67
Reading comprehension	1.234	1.185	1.523	1.171
Phonological awareness	1.738	1.813	1.883	1.914
Writing	5.898	5.894	6.225	5.905
Mathematics	0.588	0.575	0.634	0.607
English	3.024	3.006	3.649	2.936
Composite score (SD)	-0.0451	-0.0483	0.175	-0.0553

The next three figures present further descriptive evidence of the differences in achievement between the Intervention 2 group ("Coaching") and the control group. Figure 18 shows the percentage of children achieving above particular thresholds of words correct per minute, separately for the two groups of children. The scores at the 25th percentile, median (50th percentile) and 75th percentile of the distribution for the full sample of 230 schools are also indicated on the graph. In both groups there were roughly 85% of children who could read at least one word correctly. There were also similar percentages of children who managed to read all 50 words correctly within a minute (about 6%). However, throughout the range between zero and 100% there were consistently more Intervention 2 children able to surpass particular thresholds. Between the thresholds of about 10 words per minute and 25 words per minute there were consistently about 10% more children in the Intervention 2 group able to read at least that number of words than in the control group. The pattern in this graph points to the possibility that the impact of the coaching intervention was largest for children in the mid-range of the performance distribution.

Figure 18: Word recognition for Intervention 2 and Control

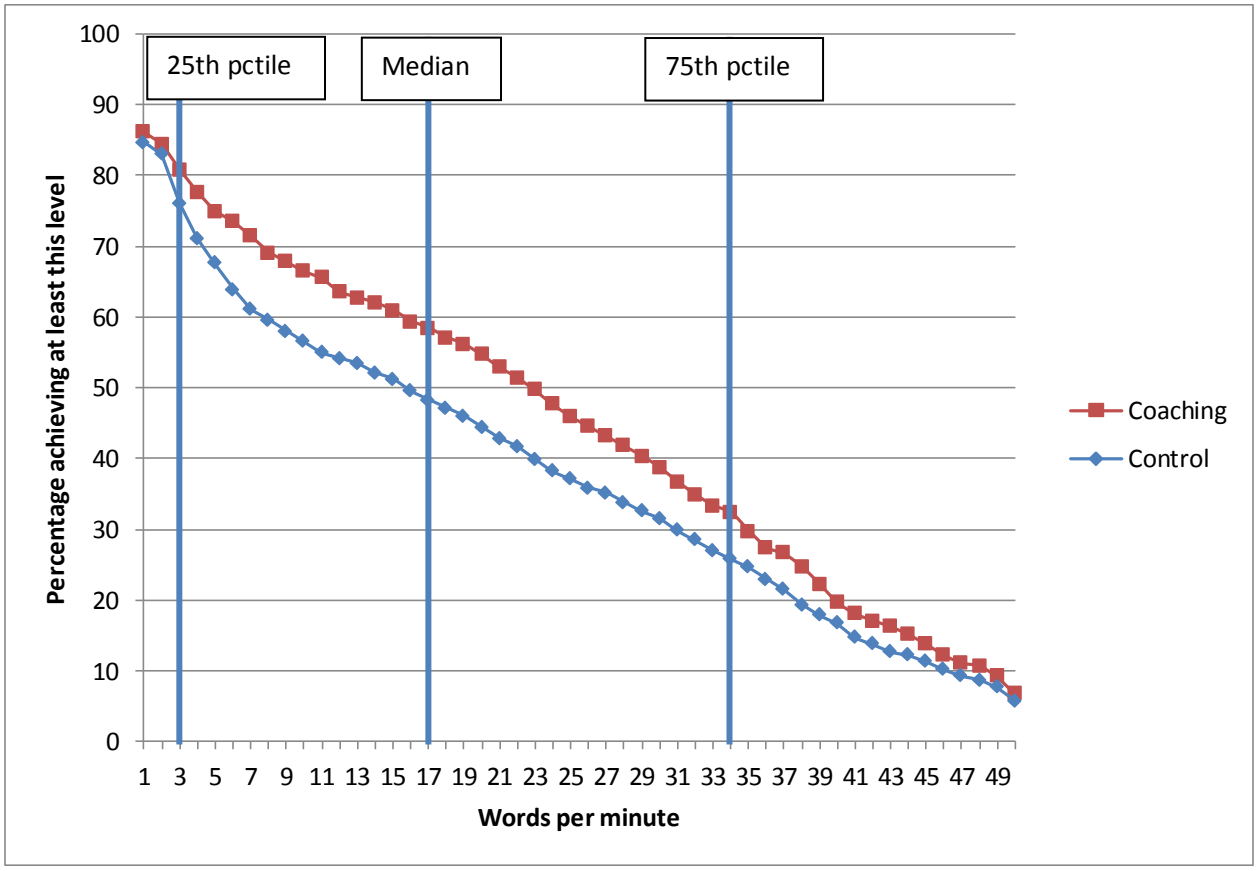


Figure 19 shows the same type of graph as above, now applied to the Paragraph reading test, which provides a measure of oral reading fluency. As before, similar proportions of children could complete reading the entire paragraph within a minute (about 10%). A significant floor effect on this item (36% of children scoring zero) means that it is difficult to say anything about the impact on the bottom end of the distribution, but we can say that whereas only about 61% of children in the control group could read at least one word correctly, there were about 72% of children in the "coaching" group who could do so. Similarly, about 10% more children surpassed the median level of achievement (23 words per minute) in the coaching group compared to the control group. Only for the top 30% of the distribution did the magnitude of this impact drop off.

Figure 19: Oral Reading Fluency for Intervention 2 and Control

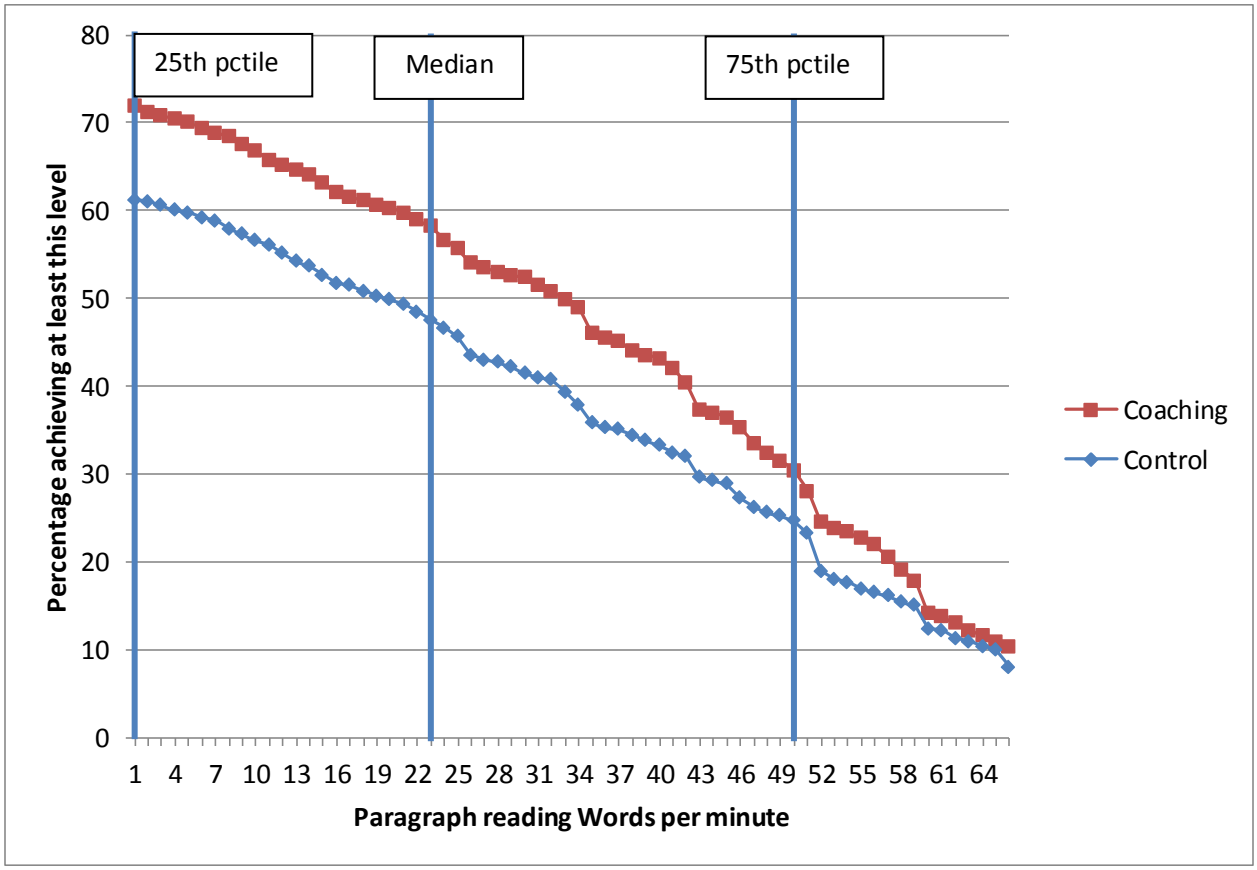
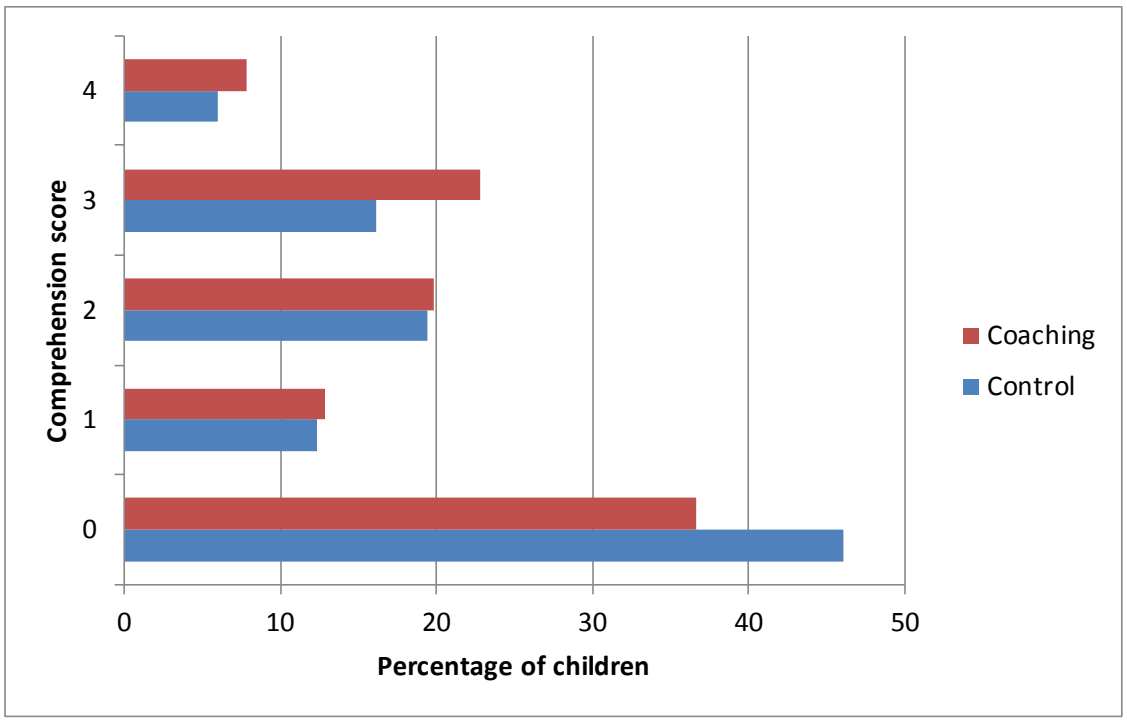


Figure 20 shows the percentages of children achieving each possible score out of 4 on the comprehension test, which was administered after the paragraph reading. Whereas 46% of children in the control group scored zero, only 37% of those in the coaching group scored zero. The entire distribution shifted upwards for the intervention group, with about 10% more children scoring either 3 or 4 out of 4 than in the control group.

Figure 20: Comprehension scores for Intervention 2 and Control



The descriptive analysis of the shifts in performance for the "Coaching" group is useful to get a sense of the magnitudes of the effects that will be described in the forthcoming regression analysis. In light of the fact that this intervention is intended to shift teaching practice and learning outcomes at a large scale, the magnitudes of the shifts in learning outcomes seen in the graphs above, though not miraculous, do appear substantial enough to warrant consideration for policy scale-up.

TREATMENT EFFECTS RELATIVE TO A YEAR OF LEARNING

As will be seen in the forthcoming sections, which estimate treatment effects using multivariate regression models, it can be difficult to gain a sense of what the effect sizes actually mean educationally. When effect sizes are reported in terms of standard deviations of test scores, this has some advantages, but the educational significance of such measurement units are not necessarily transparent, especially to audiences not versed in statistics.

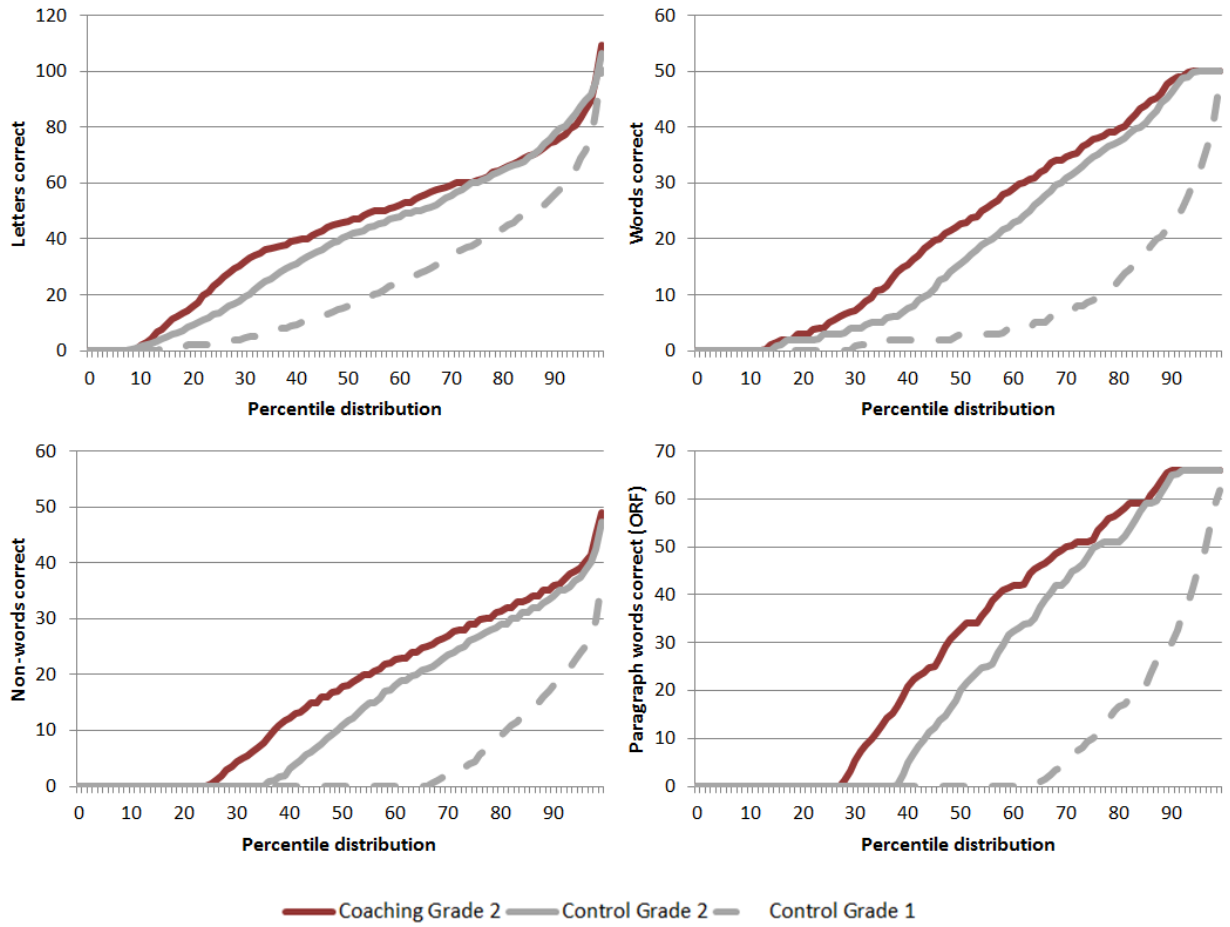
One approach is to estimate what an effect size is relative to the equivalent of 1 year of learning in the control group (which represents how much learning happens in the absence of the special intervention). In order to do this we use only those items which were common across both Waves 2 and 3. The extent to which learners in the control group improved on these items between the end of grade 1 (Wave 2) and the end of grade 2 (Wave 3) provides an estimate of 1 year of learning. The idea is then to see how much extra learning the intervention group achieved and express this additional amount as a percentage of 1 year of learning in the control group.

However, this exercise is not as simple as it may sound, as a consideration of Figure 21 demonstrates. The figure shows percentile plots (the score at each point in the percentile distribution of performance) for those items common to Waves 2 and 3. The gap between the line for the control group at the end of grade 1 and the line for the control group at the end of grade 2 represents a "year of learning". The gap between the control group at the end of grade 2 and the "Coaching" group at the end of grade 2 represents the impact of the intervention, i.e. the additional learning over and above the counterfactual. For example, one year of learning (at the 50th percentile) in terms of letter recognition is estimated to be 25 letters (the gap between 15 letters at the end of grade 1 and 40 letters at the end of grade 2). At the 50th percentile of grade 2 learners in the coaching group, the score was 46 letters correct, which was 6 letters more than the control group. Therefore, it could be estimated that the impact on letter recognition was 24% of a year of learning (6/25*100) – at least at the 50th percentile.

The graphs immediately make clear that the estimated impact expressed relative to a year of learning depends hugely on the point in the performance distribution one chooses. In particular, floor effects complicate things. For example, the impact on paragraph reading (Oral Reading Fluency) at the 25th percentile would be zero, since even in the intervention group there were more than 25% of learners who could not read a single word. Yet at the 30th percentile the impact would now be infinitely large since those in the coaching group at this point in the rank distribution were able to read 5 words, but in the control group more than 30% of learners could not read a single word.

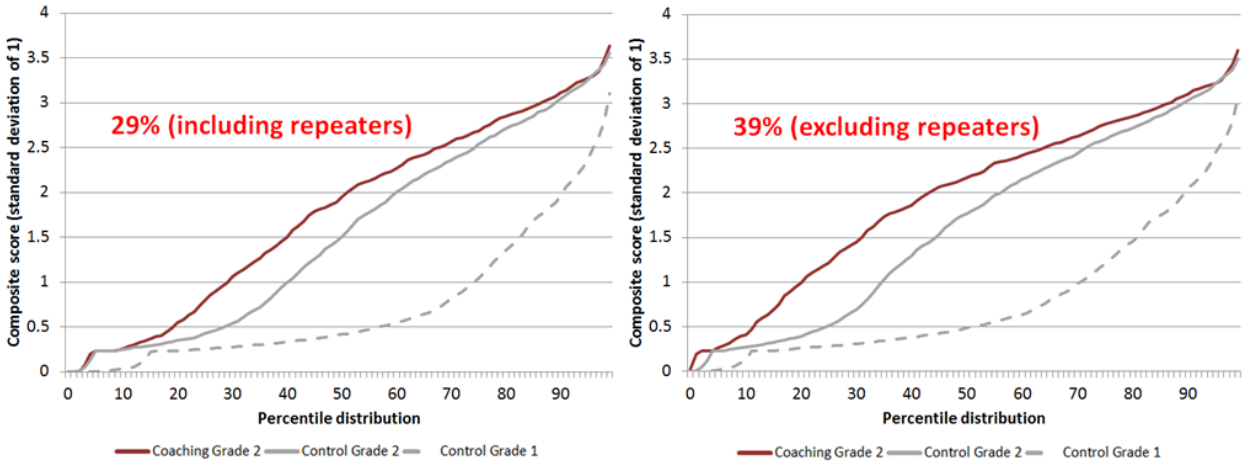
In light of these interpretation challenges, perhaps the best approach is to calculate the entire area between the control and intervention lines for grade 2 and express this as a percentage of the entire area between control at grade 1 and control at grade 2. Using this approach, we estimate that the impact of coaching on letter recognition was 24% of a year of learning, the impact on word recognition was 28% of a year of learning, the impact on non-word recognition (decoding) was 35% of a year of learning, and the impact on Oral Reading Fluency was 32% of a year of learning.

Figure 21: Percentile plots for sub-tests common across Waves 2 and 3



Another approach we adopted to estimate impact sizes relative to a year of learning, was to pool the data for Waves 2 and 3, keep only the common items, and then rerun the Principal Components Analysis to obtain composite scores that are comparable across Waves 2 and 3. For this approach we also included two writing items which were common across the two waves. After deriving these composite scores we produced percentile plots, yielding the "year of learning" graphs below. After calculating the ratio of the area between the "coaching" group and the control group at the end of Year 2 relative to the area between the control group at the end of Year 1 and the control group at the end of Year 2, the result was that we estimate the impact of being in the "coaching" intervention to be 29% of a year of learning. If we exclude those repeating grade 1 from the calculation (since they were only exposed to the intervention in Year 1), we estimate the impact to be 39% of a year of learning for those who were exposed to two years of the intervention.

Figure 22: "Year of learning" graphs based on a composite score derived from common items across Waves 2 and 3



REGRESSION ANALYSIS OF MAIN INTERVENTION IMPACTS

The descriptive results already indicate that Intervention 2 (Coaching) had the clearest effect on home language literacy and reading outcomes. Simple comparisons of outcomes across the treatment groups should not be biased thanks to the random allocation to treatment. However, regression modeling provides a more formal method for evaluating the impacts on the reading outcomes, allows for one to control for any incidental prior differences (which are bound to exist due to the limited sample size) and has the advantage of slightly reducing the size of standard errors (i.e. better precision) as more variation in the outcome can be accounted for by control variables. Table 27 shows the results of an Ordinary Least Squares regression model controlling only for the stratification dummies. Since this model has no other control variables we characterize it as a "vanilla model". This confirms the descriptive analysis above that at Intervention 2 learners were performing noticeably better than control group learners (as well as better than Intervention 1 and 3 learners) at the end of two years of interventions.

Table 27: "Vanilla" regression models of intervention impact including repeaters

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	-0.001		
	(0.083)		
Intervention 2		0.221**	
		(0.085)	
Intervention 3			-0.012
			(0.082)
Constant	-0.225*	-0.168	-0.125
	(0.122)	(0.115)	(0.127)
Observations	2,121	2,140	2,140
R-squared	0.015	0.027	0.027

Notes: Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

It is relevant to consider the differences in outcomes when excluding all learners who were repeating Grade 1 in 2016, since these learners effectively received a lower dosage of treatment. This is because in Year 2 the interventions were only conducted amongst Grade 2 teachers. Although some of the grade 1 teachers in 2016 would have benefited from the interventions in 2015, it is quite possible that the effect on their teaching would have been somewhat diminished due to no longer receiving support and not receiving new lesson plans and materials in 2016. The next table shows the same "vanilla" models but excluding repeaters. As expected, the estimated treatment effect for Intervention 2 is now larger at 0.31 standard deviations.

Table 28: "Vanilla" regression models of intervention impact excluding repeaters

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.052		
	(0.090)		
Intervention 2		0.309***	
		(0.089)	
Intervention 3			0.006
			(0.091)
Constant	-0.103	-0.098	-0.008
	(0.133)	(0.118)	(0.135)
Observations	1,758	1,772	1,781
R-squared	0.016	0.041	0.023

Notes: Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Controlling at least for baseline scores does seem appropriate given that there was a slight lack of balance at baseline with Intervention 1 learners performing a bit lower than the control group. In any case, it is standard practice to control for baseline scores in the estimation equation for the sake of improving precision. Table 29 shows the results of regression models, which control only for the stratification dummies and the baseline scores of learners. Baseline scores for each sub-test have been entered as separate control variables since this maximizes the explanatory power compared to using a baseline composite score as a single control variable. The estimated treatment effects for Intervention 2 (coaching) are virtually identical to that in the Vanilla models, whether one includes or excludes repeaters. Controlling for baseline scores does however change the estimated treatment effects for Interventions 1 and 3. Instead of being very close to zero, the estimated effects are now 0.05 and 0.06 standard deviations, respectively. When excluding repeaters, the effect for Intervention 1 is now estimated to be 0.09 standard deviations. Yet in all of these models we can still not rule out the possibility of a zero impact of Interventions 1 and 3 with even a 90% level of confidence.

Table 29: Regression models controlling for baseline scores and including repeaters

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.0545		
	(0.0788)		
Intervention 2		0.221***	
		(0.0803)	
Intervention 3			0.0606
			(0.0769)
Constant	-0.653***	-0.825***	-0.700***
	(0.201)	(0.264)	(0.226)
Observations	2,121	2,140	2,140
R-squared	0.097	0.113	0.126

Notes: Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 30: Regression models controlling for baseline scores and excluding repeaters

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.0915		
	(0.0848)		
Intervention 2		0.305***	
		(0.0837)	
Intervention 3			0.0609
			(0.0832)
Constant	-0.288	-0.564**	-0.277
	(0.235)	(0.276)	(0.269)
Observations	1,758	1,772	1,781
R-squared	0.099	0.122	0.124

Notes: Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 31 presents the results from our preferred model specification, which controls for baseline scores, district (schools are spread randomly across two districts), school mean score in the Annual National Assessments of 2014 (the most recent standardized school assessment), learner gender, parent education (according to the parent/guardian questionnaire), and two community-level controls obtained from the national census of 2011, namely a community wealth index derived from several questions about household possessions and the proportion of 13 to 18 year-olds in the community that are attending and educational institution. The motivation for including these controls is to account for any incidental differences that may exist between the treatment groups as well as to improve the precision of the estimates by increasing the explanatory power of the model. The standard errors in Table 31 are indeed slightly smaller than those in Table 29.

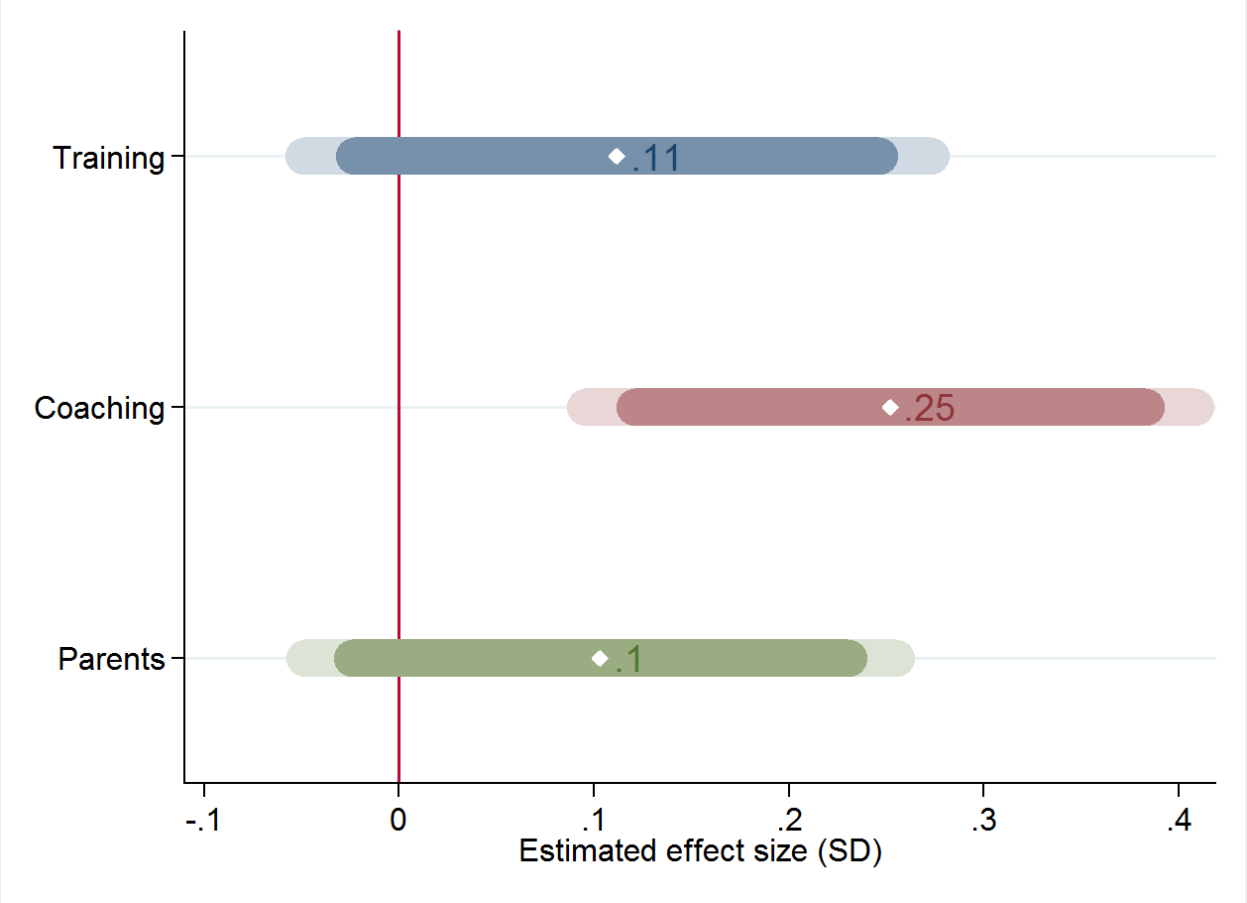
After including the additional set of controls, the estimated treatment effects are all slightly higher than in the less fully specified models. In Table 31, which includes repeaters, the coefficients on the dummies for Interventions 1 and 3 are still not statistically significant. The estimated treatment effects when including those who repeated grade 1 in 2016 are graphically represented in Figure 23.

Table 31: Year 1 regression models with full controls (including repeaters)

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.112		
	(0.0814)		
Intervention 2		0.252***	
		(0.0792)	
Intervention 3			0.103
			(0.0768)
Constant	-1.601**	-1.596**	-1.066*
	(0.624)	(0.674)	(0.591)
Observations	2,121	2,140	2,140
R-squared	0.170	0.178	0.183

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 23: Graphical representation of estimated treatment effects including repeaters showing 90% and95% confidence intervals



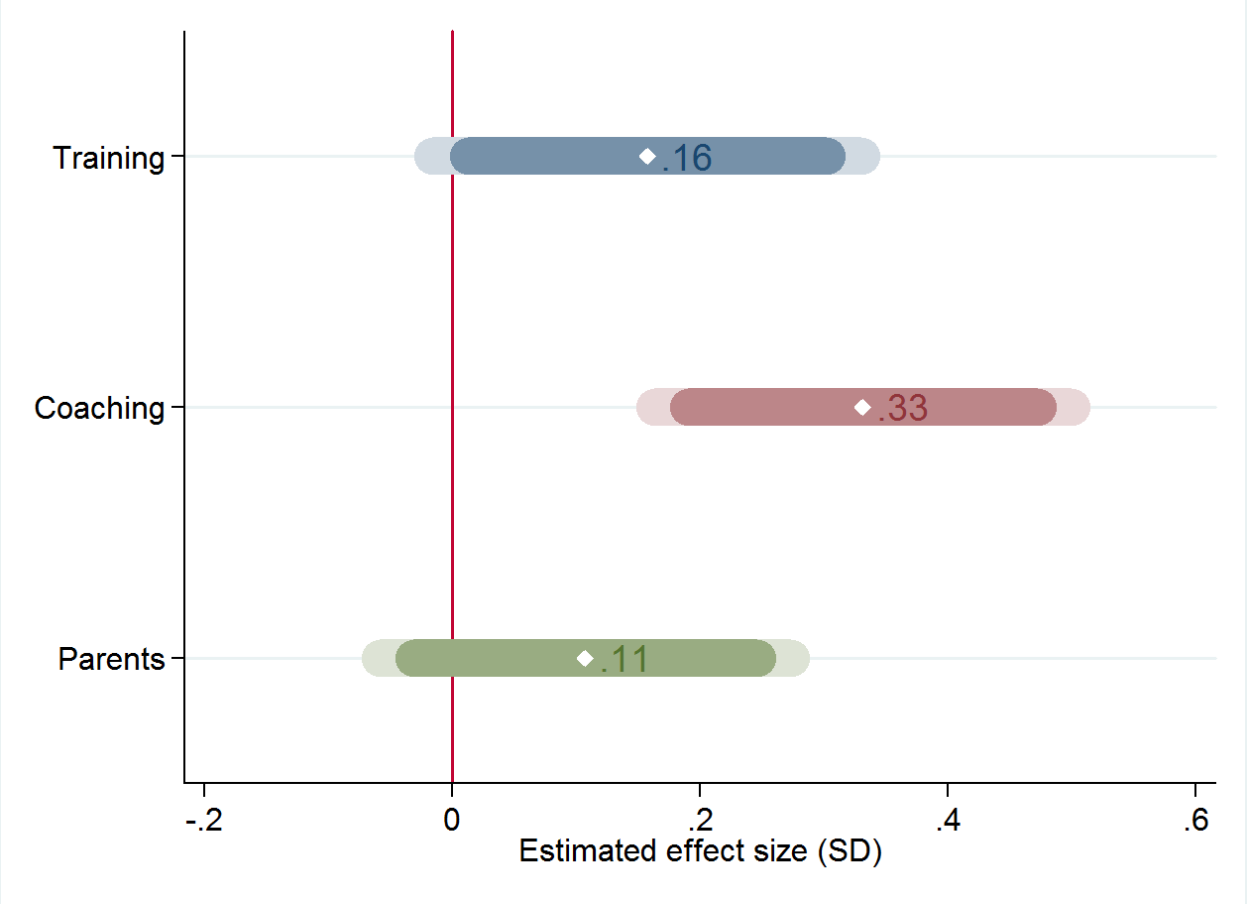
When excluding repeaters the estimated impacts of Interventions 1 and 2 increases, as one might expect given that only grade 2 teachers received additional support in 2016. We can now be 90% sure that Intervention 1 (training) had a non-zero positive impact. However, the effect size is perhaps somewhat on the small side given that this was after two years of intervention, and we cannot be 95% sure that the impact was not zero. The effect of Intervention 2 (coaching) on those children who received two years of the programme is estimated to be 0.33 standard deviations, and we can be 99% sure that there was a non-zero impact. The 90% confidence interval ranges from about 0.2 standard deviations to about 0.5 standard deviations. Relative to the RCT literature on educational interventions, this is certainly a substantial effect size. As described in an earlier section, this impact might be thought of as representing an additional 40% of a year of learning (relative to the amount of learning that occurred over a year in the control group).

Table 32: Year 1 regression models with full controls (excluding repeaters)

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.158*		
	(0.0875)		
Intervention 2		0.332***	
		(0.0853)	
Intervention 3			0.108
			(0.0840)
Constant	-1.276*	-1.450**	-0.775
	(0.695)	(0.695)	(0.680)
Observations	1,758	1,772	1,781
R-squared	0.163	0.180	0.178

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 24: Graphical representation of estimated treatment effects excluding repeaters showing 90% and95% confidence intervals



INTERVENTION EFFECTS ON SUB-TESTS

In much of the literature on measuring early grade reading, composite scores are not derived or reported. This is partly due to limited conceptual meaning of a composite score and partly because the various components or stages of learning to read are considered in relation to each other. In particular, Oral Reading Fluency and Reading Comprehension are viewed as key outcomes, while skills such as letter recognition and phonological awareness are viewed as necessary components in learning to read. We chose to derive a composite reading score for the sake of avoiding multiple outcomes for every type of analysis reported in our analysis. However, we also present the main regression results for each of the sub-tests, including the English and mathematics items.

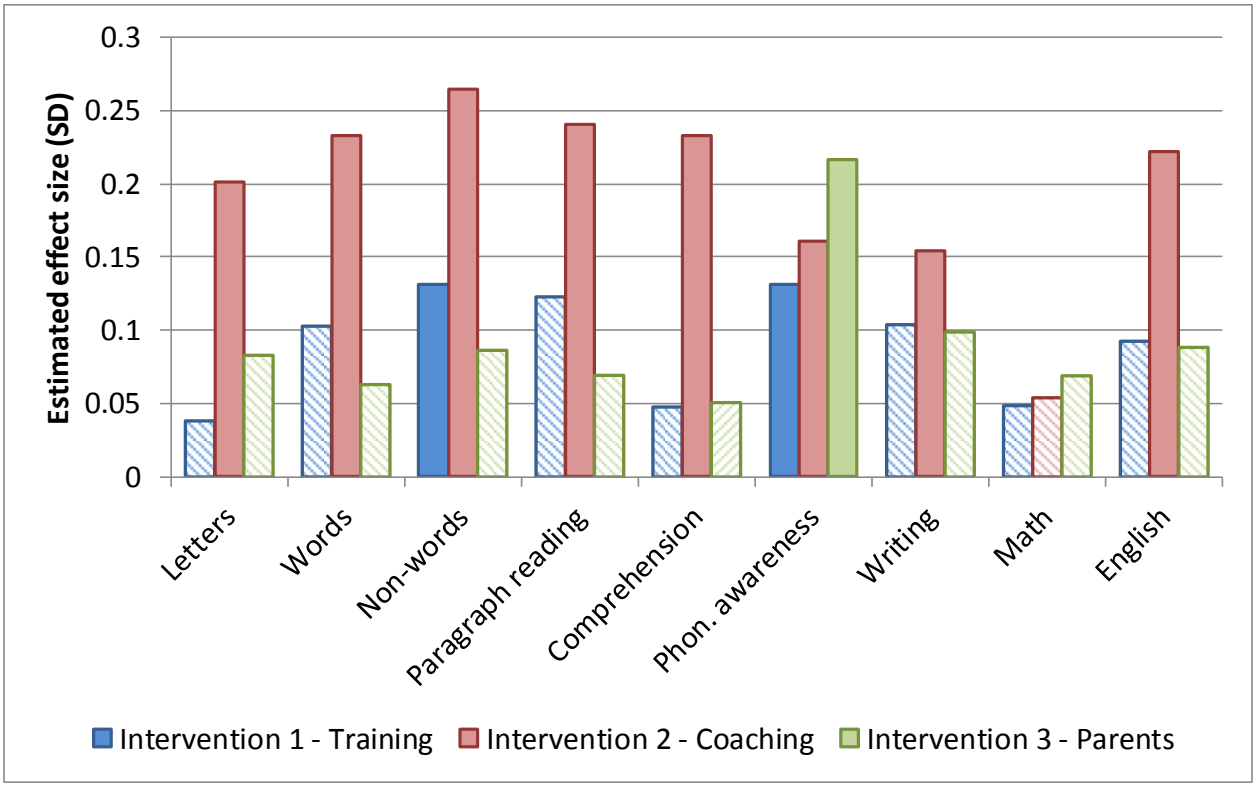
The same set of controls as used in the main specification above was included in these models. Figure 25 graphically presents the results of regressions for each of the sub-tests, with effect sizes expressed in terms of standard deviations. Solid bars represent statistical significance at

the 90% level. Several results are worth highlighting. Firstly, there were no negative effects of any intervention on any sub-test. The training intervention appears to have had moderate positive effects on most sub-tests with statistically significant impacts on non-word recognition, paragraph reading (Oral Reading Fluency) and phonological awareness. However, the effect sizes on word recognition and writing were not much smaller. Therefore, it seems unwarranted to make strong conclusions about Intervention 1 being particularly effective or ineffective in specific dimensions of reading literacy.

Intervention 2 registered statistically significant positive effects on all home language sub-tests, with similar effect sizes across the sub-tests. Therefore, it is not as if there is any one or two dimensions of learning that are driving the positive results for Intervention 2. There was no statistically significant effect of Intervention 2 on the rudimentary mathematics test. This means that we have no evidence of a negative effect through crowding out of teaching time for mathematics. Interestingly, we observe a positive effect on English (significant at the 95% level). Here too, there could have been a negative crowding out effect especially since the national curriculum gives teachers in the Foundation Phase the choice to either spend 3 hours a week on English and 7 hours on home language literacy or 2 hours on English and 8 hours on home language literacy. In our intervention, lesson plans were designed on the assumption that teachers would opt for the full 8 hours dedicated to home language literacy. The positive effect on English could be attributable to improved underlying language ability (as obtained through the home language intervention) or simply due to improved classroom management and transferable instructional methods acquired through the coaching intervention. Either way, this is a highly encouraging finding for the intervention.

Although the overall impact of the parent intervention was small and statistically insignificant, it does appear to have had a significant (at the 95% level) positive impact on phonological awareness. It is possible that this is a false positive – 1 out of 20 zero effects could be expected to come out as a false positive at the 95% level, and here we have one out of 9 outcomes for the parent intervention being positive. However, phonological awareness was certainly the component of the learner test that was most directly targeted through the parent meetings. Sound games were a key method taught to parents to use at homes in the development of their child's phonological awareness.

Figure 25: Effects on sub-tests



INTERVENTION EFFECTS ON SUB-GROUPS OF INTEREST

The question of whether an intervention had a differential effect on various sub-groups is important for policy and for understanding when and how these interventions are effective. Therefore, we collected a considerable amount of contextual information about learners, their teachers and the schools they are in. When an intervention has a differential effect on various subgroups this is often referred to in the literature as a heterogeneous treatment effect. However, there is a risk when investigating numerous possible heterogeneities of so-called data mining – that sooner or later a statistically significant result is bound to occur. The existence of a midline assessment as well as an endline assessment reduces this risk somewhat. We were particularly cautious in interpreting observed heterogeneous treatment effects in the Midline report. However, to the extent that we observe similar heterogeneities in the midline data and again after the second year of interventions we can be more confident that a genuine effect is occurring.

DIFFERENTIAL INTERVENTION EFFECTS BASED ON LEARNER CHARACTERISTICS

As laid out in our Pre-Analysis Plan, we investigate learner-level treatment heterogeneity based on learner gender, learner age, and the initial performance of the learner at the start of grade 1. In general, we investigated these for all three interventions. However, when the results prompted us to dig a little deeper, the focus was mainly on Intervention 2, since this is where the clearest impact was observed. Table 33 shows summary statistics for the various learner-level variables for which we present estimates of heterogeneous effects.

Table 33: Summary statistics for learner-level variables used in analysis of heterogeneous effects

	count	mean	min	p25	p50	p75	max
Baseline composite score	4538	0	-1.833	-0.580	-0.294	0.268	5.403
Female (dummy)	4538	0.46	0	0	0	1	1
Learner age (years)	4512	6.51	4	6	6.417	6.833	10.5

Table 34 presents the regression results for models which measure the interaction effect of gender and treatment. The main effect for each intervention is here interpreted as the effect of the intervention for boys, which the coefficient on the interaction term represents the additional effect (whether positive or negative) for females. The results indicate that each of the interventions had a positive statistically significant impact for boys, but a lesser effect for girls. For Interventions 1 and 2, this is consistent with what was observed after one year of interventions. Therefore, it seems fair to conclude that the two pedagogical interventions, especially the “coaching” intervention, are helping boys catch up some of the way to girls.

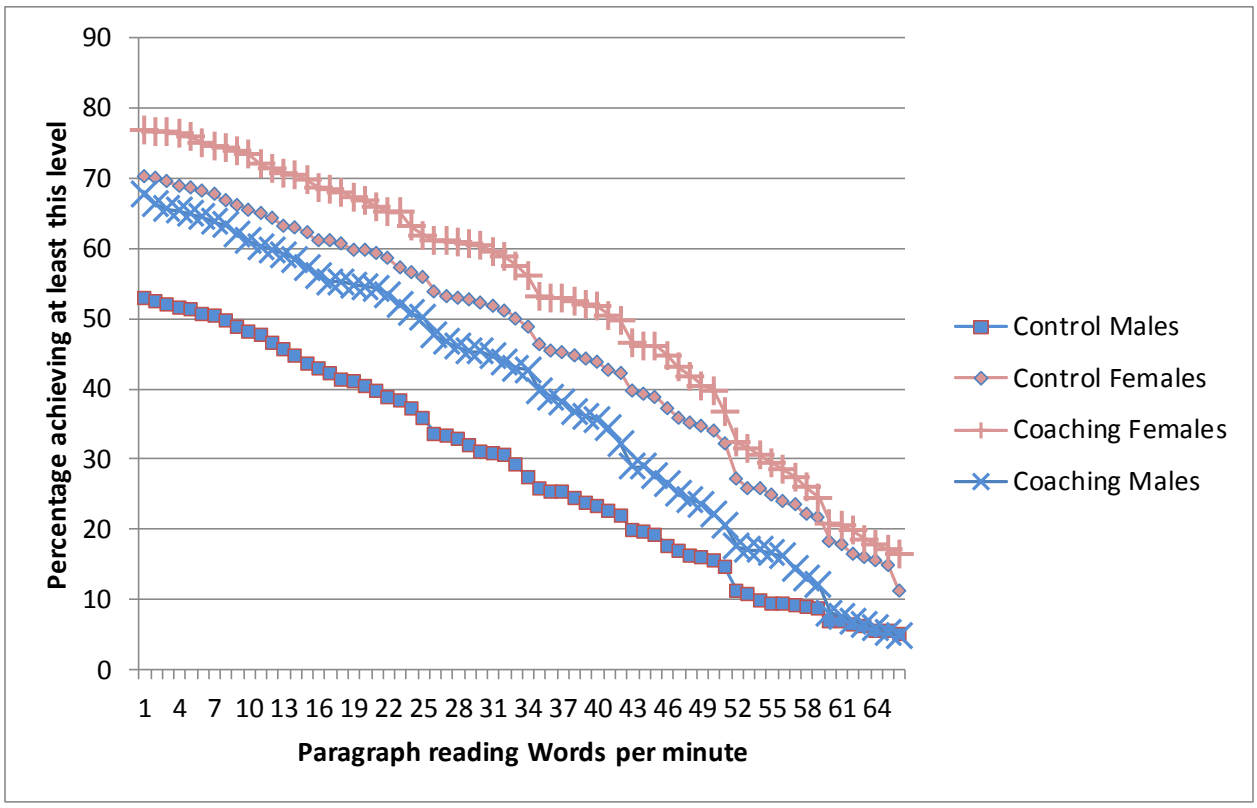
Table 34: Intervention effects by learner gender

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.165*		
	(0.0881)		
Female_X_ Intervention 2	-0.116		
	(0.0852)		
Intervention 2		0.307***	
		(0.0868)	
Female_X_ Intervention 2		-0.117	
		(0.0905)	
Intervention 3			0.147*
			(0.0833)
Female_X_ Intervention 3			-0.0955
			(0.0783)
Constant	-1.607**	-1.620**	-1.082*
	(0.624)	(0.672)	(0.590)
Observations	2,121	2,140	2,140
R-squared	0.171	0.179	0.184

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 26 shows a type of inverse cumulative distribution function – it indicates the percentage of children achieving at least at a certain level, in this case on paragraph reading (or Oral Reading Fluency) at the end of Year 2. The graph shows this for boys and girls in the control group and in the "coaching" group. In both groups of schools, girls are outperforming boys, yet the gap has narrowed for the treatment group. For example, about 30% of girls in the control group could not read a single word compared to nearly 50% of boys. In the "coaching" group, however, the percentage of boys who could not read even one word was just over 30% (not far off where girls are in the control group), while nearly 80% of girls could read at least one word. The reasons why this structured learning programme may be helping boys catch up to girls are difficult to identify with certainty. However, based on some of the changes we are observing in classroom practice (to be presented in a later section), it seems reasonable to suggest that the improved classroom management and increased individualized attention may assist boys to be engaged in learning activities. In contexts of large classes it seems especially likely that boys may be less engaged in active learning. Machin and McNally (2005: 363) note evidence from the United Kingdom that boys benefit from highly structured methods of teaching and are more likely than girls to "respond negatively to poor teaching through disengagement and indifference or through disruptive behaviour".

Figure 26: Oral Reading Fluency by gender and treatment group



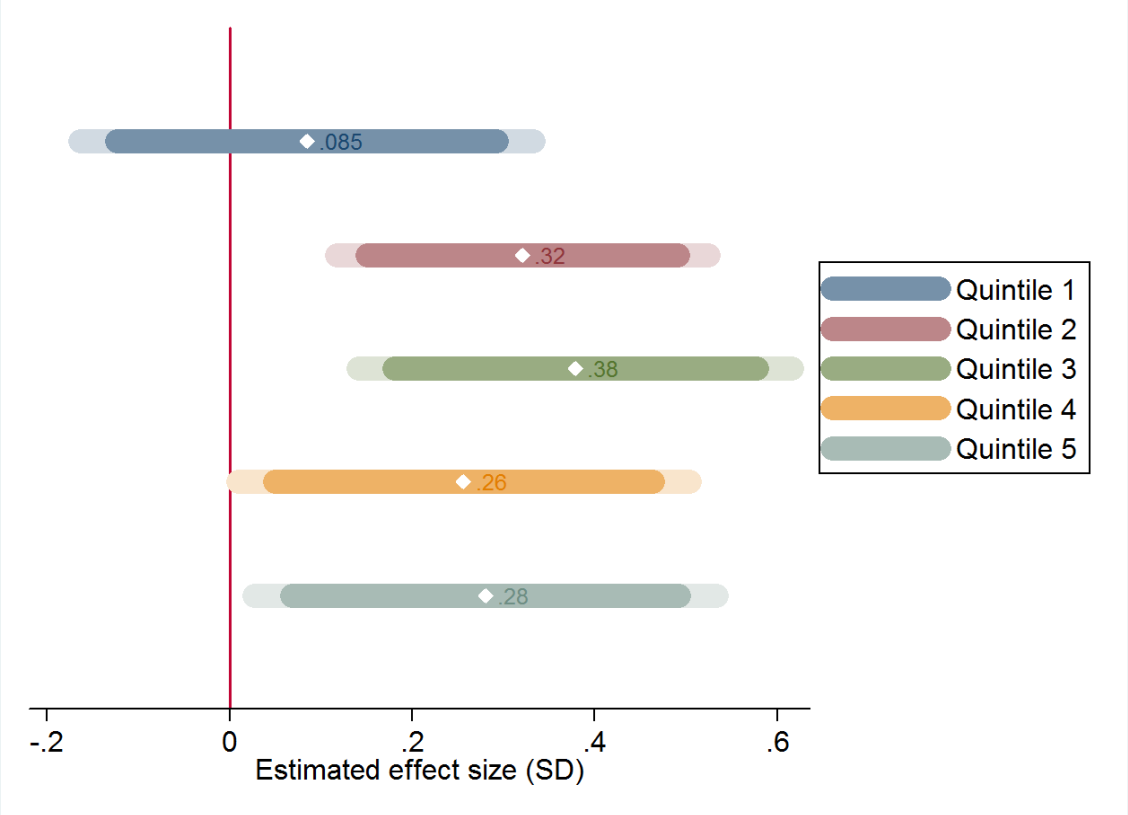
We also investigated whether the effect of the interventions depended on the age of the learner. Table 35 shows the regression coefficients from models where the age of the learner is interacted with the intervention group. The negative (though not statistically significant) coefficients for the main treatment effect indicate that for the very youngest learners there may not have been an impact. However, none of the interaction terms come through significantly either, suggesting that there is no clear story of a differential effect by age. To investigate this further, we ran another model (not shown here) in which age and it's interaction term were allowed to take on a quadratic functional form, instead of a linear functional form as in Table 35. There was some indication that the "coaching" intervention had a larger treatment effect for those in the mid-range of the age distribution, though again this result was not statistically significant. A final way of investigating this was to split the sample into 5 quintiles of learner age and run five separate regressions. This analysis was only done for the coaching intervention. The results are reported graphically in Figure 27. This also suggests larger effects for those in the mid-range of the age distribution, who are also roughly of the appropriate age for grade. This provides tentative evidence that improvements to the enactment of the curriculum are most effective when children are of the appropriate age for grade, and thus confirms the importance of adherence to age for grade policies.

Table 35: Intervention effects by learner age

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
T1	-0.0567		
	(0.403)		
Age_X_Intervention 1	0.0263		
	(0.0612)		
T2		-0.0889	
		(0.480)	
Age_X_Intervention 2		0.0528	
		(0.0736)	
T3			-0.186
			(0.456)
Age_X_Intervention 3			0.0443
			(0.0675)
Constant	-1.643**	-1.635**	-1.003
	(0.724)	(0.737)	(0.689)
Observations	2,113	2,135	2,133
R-squared	0.172	0.178	0.182

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 27: Effect of coaching intervention by quintile of learner age



The next important question with respect to differential treatment effects is whether the impact was different for stronger or weaker learners. If the impact of an intervention is larger for weaker learners this can be seen as an equity-enhancing programme. Nevertheless, the sample for this project has been restricted to non-fee paying schools, which serve about 70% of South Africa's learners. So the most affluent part of the South African population is not included in this analysis.

There are two main ways we investigate the question of whether the impact was different across the learner performance distribution. The first is to investigate whether the impact depended on baseline learner performance: Did initially weaker/stronger learners benefit more from the intervention? The second approach is to observe the performance distributions for treatment and control groups at the end of two years of interventions and see whether the differences in performance vary across the distribution. One would not expect the two approaches to yield completely opposing results, although the calculation is different leading to the possibility of such. Specifically, if there is a lot of rank mobility during the course of the intervention (initially weaker learners moving up the rank distribution and vice versa) but little difference in the overall shape of the distribution this would affect the first approach but not the second. From a policy point of view the second approach is arguably more relevant since the overall level of inequality is perhaps more important than *who* is moving up and down the distribution.[10]

The first approach can be followed by interacting the treatment variable with baseline learner scores. Table 36 shows the results of this approach for a regression in which the relationship between treatment impact and baseline score is assumed to be linear. For all three interventions similar overall effect sizes of the interventions are obtained as compared to the main results discussed earlier. The interaction terms are not significantly different to zero, indicating that there was no obvious linear relationship between baseline score and the impact of any intervention.

[10] An exception would be if historically unjust patterns of inequality are being perpetuated, which would of course be a concern to the policymaker.

Table 36: Intervention effects by learner baseline scores (linear models)

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.111		
	(0.0811)		
BL_X_Intervention 1	-0.00529		
	(0.0788)		
Intervention 2		0.249***	
		(0.0793)	
BL_X_Intervention 2		0.0426	
		(0.0825)	
Intervention 3			0.107
			(0.0762)
BL_X_Intervention 3			0.101
			(0.0649)
Constant	-1.604**	-1.614**	-1.038*
	(0.635)	(0.669)	(0.595)
Observations	2,121	2,140	2,140
R-squared	0.170	0.178	0.186

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

It is quite possible, however, that a non-linear relationship may exist (for example, where those in the middle of the distribution benefitted more than the initially weakest and strongest learners). To allow for this possibility, another model was fitted assuming a quadratic functional form. This is achieved by entering the square of the baseline score (BL2) into the regression equation as well as the interaction between treatment and the squared baseline score. For both Interventions 1 and 2, it appears that the impact increases with learner baseline score but a declining rate, flattening off and even declining slightly at the very top end of the distribution.

Table 37: Intervention effects by learner baseline scores (quadratic models)

	Intervention 1 (Training)	Intervention 2 (Coaching)	Intervention 3 (Parents)
Intervention 1	0.193**		
	(0.0898)		
BL_X_Intervention 1	0.0873		
	(0.0962)		
BL2_X_ Intervention 1	-0.101***		
	(0.0370)		
Intervention 2		0.294***	
		(0.0837)	
BL_X_Intervention 2		0.277***	
		(0.105)	
BL2_X_ Intervention 2		-0.0895***	
		(0.0324)	
Intervention 3			0.0934
			(0.0903)
BL_X_Intervention 3			0.113
			(0.109)
BL2_X_ Intervention 3			-0.0172
			(0.0325)
Constant	-1.554***	-1.305**	-0.917
	(0.590)	(0.598)	(0.564)
Observations	2,121	2,140	2,140
R-squared	0.151	0.167	0.157

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A final way to investigate and represent the differential impact of the interventions according to baseline scores, was to run separate regressions for four quartiles of baseline achievement. The results for the coaching intervention are presented graphically below. The results are in line with Table 37 above, as one would expect, but are perhaps easier to comprehend. It would appear from this method, that the impact of coaching may have been zero amongst initially weakest 25% of learners at baseline. The estimated effect then increases slightly with each subsequent quartile, with an estimated impact of 0.33 standard deviations for the initially strongest 25% of learners. It is possible, however, that part of the reason for initially weaker learners benefiting less was because they were more likely to repeat grade 1 in 2016 and thus only receive one year of the treatment. In order to assess this possibility, we reran the regression model excluding repeaters from the sample, on the basis that they did not receive the full two years of the programme. The results are presented in Figure 29. Clearly, the estimated effect sizes increase noticeably for the bottom two quartiles, where the highest proportions of repeaters are located. Although, the estimated effect of 0.15 standard deviations is still not statistically significant it is still a meaningful effect size. This would suggest that even initially weak learners did benefit from the coaching intervention, as long as they were exposed to two years of the intervention, but we cannot be 90% sure of this. Further, when excluding repeaters, the estimated effect of the intervention appears largest in the mid-range of the

distribution. But again, the confidence intervals overlap considerably between the various quartiles so we cannot conclude this with any statistical certainty.

Figure 28: Intervention effects by quartiles of baseline performance

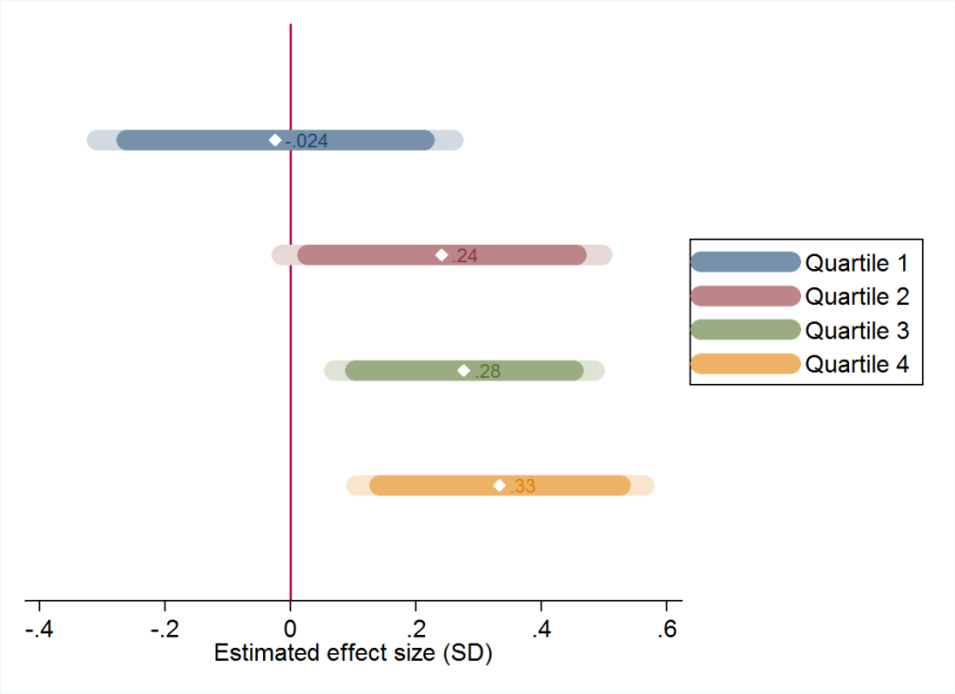
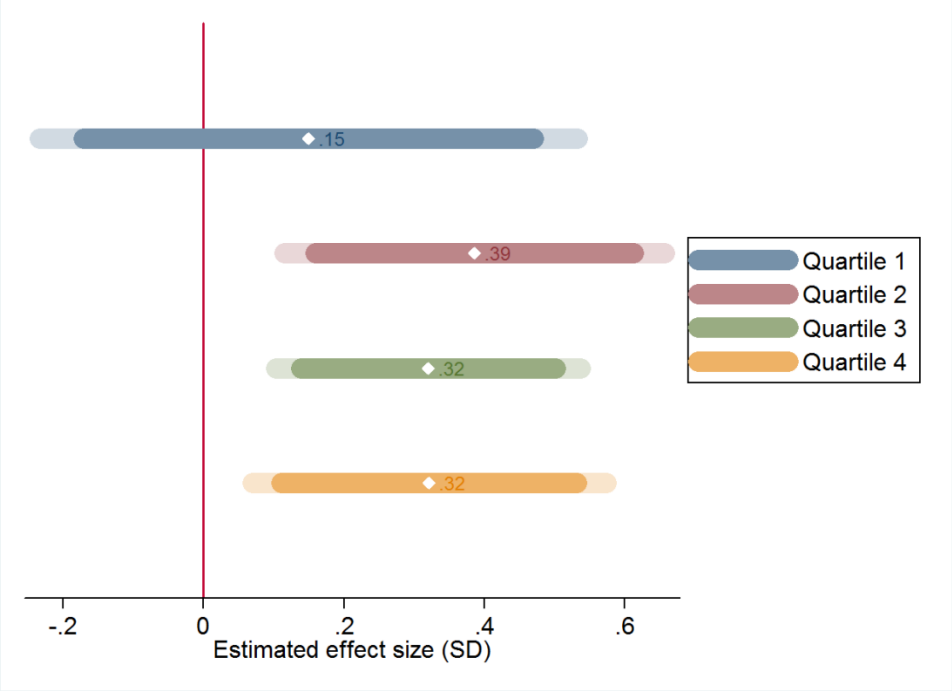


Figure 29: Intervention effects by quartiles of baseline performance (excluding repeaters)



The second approach, as discussed above, is to observe the performance distributions for treatment and control groups at the end of two years of interventions and see whether the differences in performance vary across the distribution. The method we use to do this is known as quantile regression. This estimates the effect of the intervention at various points in the distribution of the performance outcome. It asks, for example, what is the impact on the 10th percentile of performance, on the 20th percentile, on the 30th percentile, etc. We present the results of quantile regressions measuring the effect of the "coaching" intervention on the Midline (Year 1) composite test score and the Year 2 composite score in Figures 30 and 31, respectively. In both graphs the line plots the estimated effects across the performance distribution, while the shaded area represents the 95% confidence interval around the estimated effects.

The impact on both Year 1 and Year 2 scores was near zero at the 10th percentile of the distribution, and then quickly rose across the distribution, peaking at the 80th percentile in the case of the Year 1 results and the 50th percentile in the case of the Year 2 results. This confirms the earlier analysis indicating that the impact of the intervention appears to have been greatest in the middle to upper parts of the performance distribution, though not at the very top of the distribution. Importantly, there is no evidence of a negative effect for any part of the performance distribution.

One implication of this finding is that structured learning programmes, making use of lesson plans, may benefit a certain section of the performance distribution more, depending on the level at which the lessons are pitched, but at least in the case of this particular programme, no group was harmed and the level appears to have been pitched towards the middle of the learner proficiency range.

Figure 30: Quantile regression of "Coaching" Intervention impact on Year 1 scores

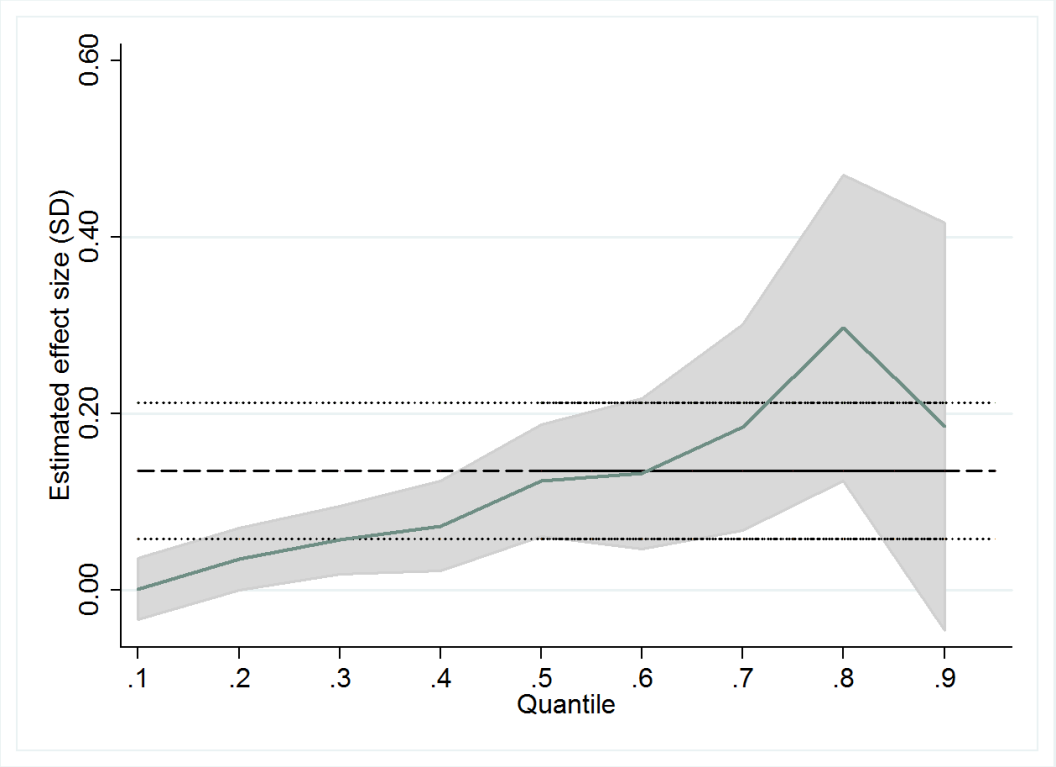
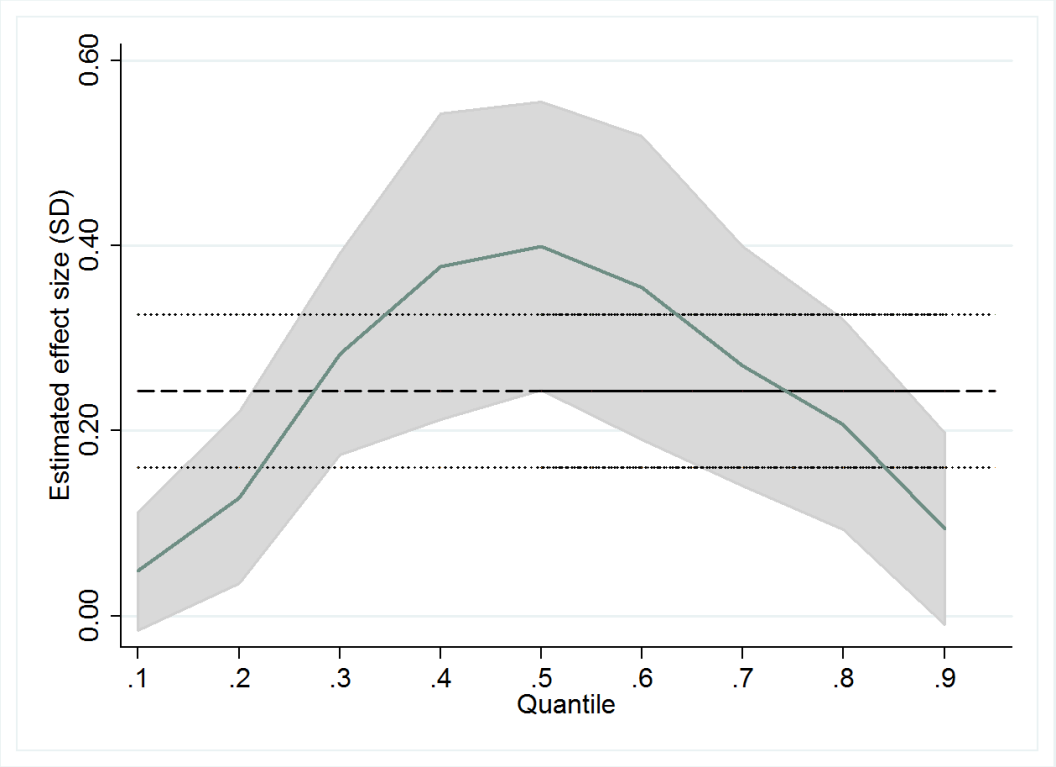


Figure 31: Quantile regression of "Coaching" Intervention impact on Year 2 scores



DIFFERENTIAL INTERVENTION EFFECTS BASED ON SCHOOL CHARACTERISTICS

A number of school-level variables might be expected to influence the effectiveness of the EGRS interventions. The academic performance level of the school may be an indication of the school's readiness for a structured pedagogic programme. School size may affect the intervention in so far as it affects the number of classes in a grade, which affects the role of peer support amongst teachers in the school. School size also interacts with the rurality of a school since deep rural schools are considerably smaller on average. The EGRS has been implemented in two education districts, and the level of district support for the programme may influence its effectiveness. The condition of school facilities may be a proxy for community level poverty or for school management quality, both factors which may influence the effectiveness of an additional school support intervention. The levels of parental education and employment are also proxies for socio-economic status, which may influence the effectiveness of, in particular, the parental involvement intervention.

Table 38 reports summary statistics for the school-level variables used in this section's subgroup analysis. The first three variables are based on the randomized stratification design of the sample and are therefore ideally suited for a balanced subgroup analysis. The sample was stratified based on school performance in the 2014 Annual National Assessments (ANA), school size, and the socio-economic status of the school as defined by the official school poverty quintile classification. Five out of the 10 strata were classified as schools with weak ANA performance, whilst four of the 10 strata consist of schools defined as being smaller than a particular enrolment threshold, and four of the 10 strata were defined as poor. 77% of schools can be classified as being located in rural settings, based on the school principal questionnaire. 23% of schools are in one education district and the remaining 77% of schools in the other. 28% of schools were described by fieldworkers as having very well maintained facilities, while the remaining schools were either in a state of disrepair or showing some signs of disrepair. The average class size in the sample was 41.58 with a considerable amount variation – 25% of classrooms, for example, had more than 48 learners. The school principal was also asked to estimate the levels of parent education and parent employment rates for the school. Nearly 60% of school principals estimated that most parents in their school have not completed secondary school. Nearly half the school principals reported that less than 20% of parents at their school were employed.

Table 38: Summary statistics for school-level variables used in sub-group analysis

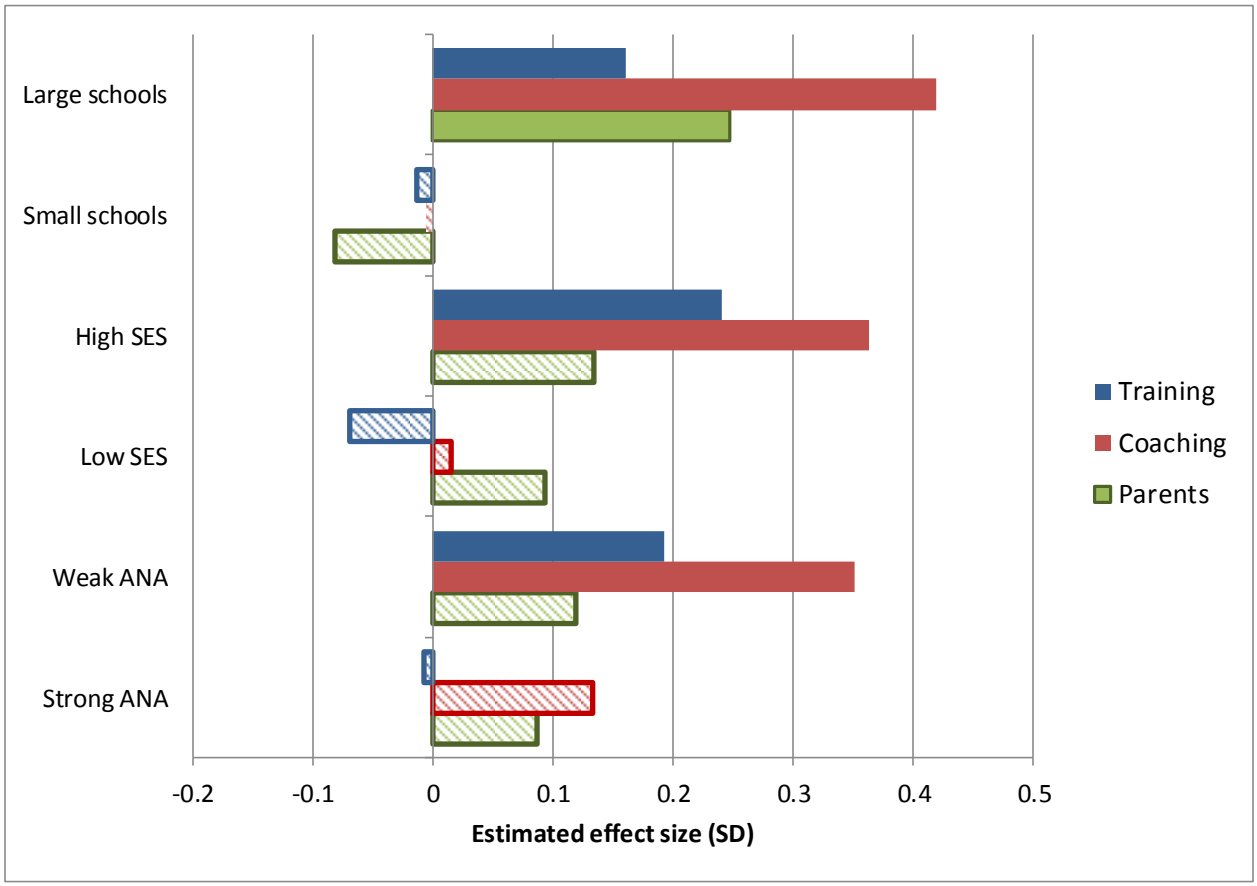
	count	mean	min	p25	p50	p75	max
Weak ANA strata	230	0.50	0	0	.5	1	1
Small school strata	230	0.40	0	0	0	1	1
Poor school strata	230	0.40	0	0	0	1	1
Rural	230	0.77	0	1	1	1	1
District A	230	0.23	0	0	0	0	1
Good facilities	228	0.28	0	0	0	1	1
Class size	219	41.58	10	35	41	48	87
Wave2_pared_low	226	0.58	0	0	1	1	1
Wave3_pared_low	226	0.57	0	0	1	1	1
Wave2_low_employ	227	0.47	0	0	0	1	1
Wave3_low_employ	225	0.46	0	0	0	1	1

Note: All variables in this table are dummy variables except for the class size variable

When running separate regressions for large schools and small schools it is apparent that none of the interventions had any significant effect within the strata consisting of smaller schools, but all three interventions had significant positive effects within the strata consisting of larger schools. Similarly, the training and coaching interventions had large statistically significant impacts within the strata consisting of relatively higher socio-economic status schools but no impacts within the strata consisting of relatively low socio-economic status schools. Somewhat unexpectedly the training and coaching interventions had clear positive impacts within the strata with weaker performance in the 2014 ANA, but no significant impact amongst schools with someone better performance in ANA. All three of these subgroup effects may well be interacting with another urban-rural subgroup effect, a result which is reported on in Table 39, where it becomes clear that no interventions had a significant positive effect within rural areas but all three had significant positive effects in the urban township settings. Urban schools are more likely to be large schools and also higher socio-economic status schools on average. It would also appear that the significant positive effects within those strata consisting of initially weaker performing schools is largely being driven by a group of initially low performing urban township schools, which experienced strong gains throughout the intervention period.

There is also a significant interaction between the education district and effectiveness of interventions. Again this may have some overlap with the urban-rural interaction since the district in which the interventions had greater effects is also the district with a higher proportion of urban township schools. No significant interaction between the condition of school facilities and the effect of interventions was observed.

Figure 32: Estimated treatment effects for various subgroups based on stratification of the sample



Note: Solid bars denote an estimated effect that is statistically significantly different to zero with at least 90% level of confidence, whilst a shaded bar denotes that we cannot be 90% sure that the effect is different from zero.

Table 39: Estimated treatment effects for various subgroups of schools

	Rural interaction	District interaction	Facilities interaction
Intervention 1	0.450***	0.0565	0.0881
	(0.165)	(0.0868)	(0.0988)
Intervention 2	0.753***	0.166**	0.158*
	(0.149)	(0.0808)	(0.0893)
Intervention 3	0.386***	0.0376	0.0579
	(0.135)	(0.0848)	(0.0892)
Rural_X_ Intervention 1	-0.391**		
	(0.182)		
Rural_X_ Intervention 2	-0.631***		
	(0.177)		
Rural_X_ Intervention 3	-0.312*		
	(0.162)		
Rural	0.316**		
	(0.139)		
District_X_ Intervention 1		0.333	
		(0.205)	
District_X_ Intervention 2		0.368*	
		(0.200)	
District_X_ Intervention 3		0.365**	
		(0.183)	
District		-0.324**	
		(0.140)	
Good facilities_X_ Intervention 1			0.0939
			(0.167)
Good facilities_X_ Intervention 2			0.281
			(0.174)
Good facilities_X_ Intervention 3			0.229
			(0.182)
Good facilities			-0.0361
			(0.120)
Constant	-2.081***	-1.581***	-1.680***
	(0.449)	(0.445)	(0.444)
Observations	3,781	3,781	3,744
R-squared	0.181	0.178	0.178

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

It is difficult to know exactly what the main causal drivers were behind the somewhat overlapping subgroup effects across urban and rural locations, education districts, school size and school socio-economic status. Whilst rural schools are poorer and smaller than urban township schools, neither of these factors fully account for the urban-rural subgroup effect. Similarly we note that both learner and teacher absenteeism is higher in rural areas, but neither does this account for the urban-rural heterogeneous treatment effect. It may be that a combination of disadvantageous factors in deep rural settings precludes interventions from having a positive impact.

Another factor which appears to be different between urban and rural areas is the extent to which schooling and coaching meetings were disrupted. Table 40 shows the percentage of schools in which coaching visits were disrupted for a variety of reasons, according to Class Act monitoring data. Factors such as difficulties in accessing schools in the rainy season, social unrest, teacher absenteeism, attending memorial services during normal school hours, choir competitions and sporting events interfering with normal teaching time, and difficulties in communicating with teachers due to poor mobile phone coverage, were all more prevalent in rural schools than in urban schools. It seems fair to conclude that if these factors commonly disrupted coaching visits they would also be likely to regularly disrupt normal teaching time in rural schools, and sufficient teaching time is probably a precondition for a structured pedagogic program following prescribed schedule to be successful.

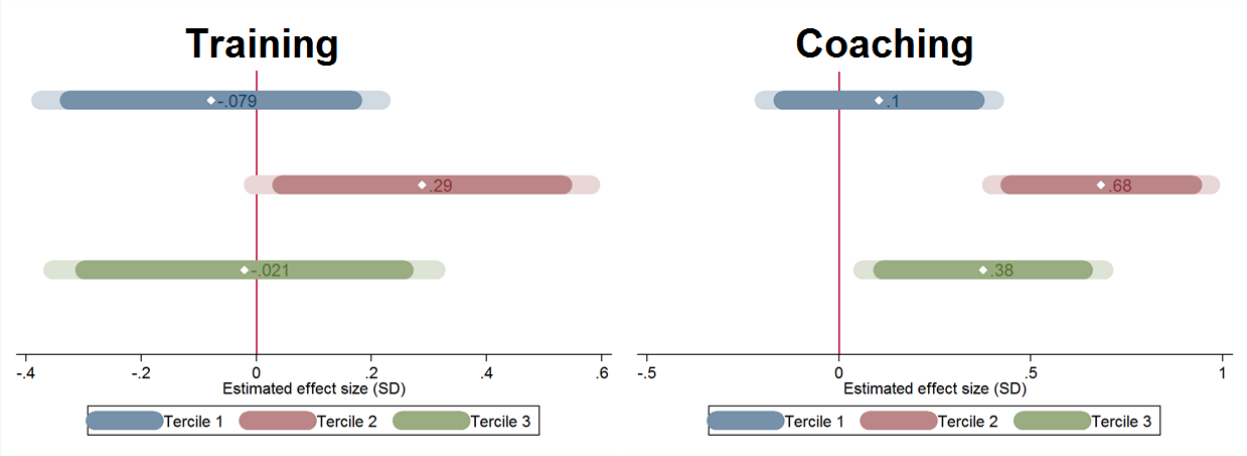
Table 40: Percentage of schools "Greatly affected" by various disruptions

	URBAN	RURAL
School access in rainy season	0%	48%
Disruptions as a result of unrest	7%	39%
Teacher absenteeism	7%	35%
Memorial services	0%	33%
Choir competitions	14%	21%
Sporting events	7%	24%
Poor cell phone coverage	0%	42%
OTHER (e.g. No electricity)	7%	21%

Source: Class Act monitoring data

We also investigated the influence of class size on treatment effects. As can be seen in Figure 33, both the teacher support interventions ("training" and "coaching") had the largest impacts in relatively large classes (38 to 45 learners). In smaller classes, it may be that teachers in the control schools are already able to effectively manage classrooms, provide structured learning and differentiated attention to a variety of learners, but in larger classes the EGRS interventions helped teachers to do so in a more challenging environment. In the very largest classes (50 plus learners), however, the EGRS interventions were somewhat less effective, possibly indicating that beyond a certain threshold it is very difficult to conduct effective teaching.

Figure 33: Treatment effects by tercile of class size



Notes: Children repeating grade 1 in 2016 are excluded from this analysis since we did not record class size in Grade 1 classes of 2016; In the cases of 157 Grade 2 learners (5% of the relevant sample) class size was missing and was imputed from the size of the other grade 2 classes in the same school. The results were negligibly different when excluding this 5% of learners.

Table 41 shows that the level of parent education and parent employment rates, as estimated by the school principal, did not seem to play a significant role in determining the effectiveness of interventions.

Table 41: Treatment heterogeneity according to Principal's assessment of parental education and employment levels

	Parent education (Wave 2)	Parent education (Wave 3)	Parent employment (Wave 2)	Parent employment (Wave 3)
Intervention 1	0.0748	0.157	0.156	0.233**
	(0.136)	(0.141)	(0.107)	(0.111)
Intervention 2	0.180	0.234*	0.246***	0.307***
	(0.113)	(0.122)	(0.0906)	(0.108)
Intervention 3	0.00946	0.139	0.184*	0.200*
	(0.114)	(0.126)	(0.102)	(0.108)
Low Par.Ed_X_ Int. 1	0.110			
	(0.174)			
Low Par.Ed_X_ Int. 2	0.142			
	(0.152)			
Low Par.Ed_X_ Int. 3	0.199			
	(0.158)			
Low Par.Ed	-0.124			
	(0.107)			
Low Par.Ed_X_ Int. 1		-0.0614		
		(0.167)		
Low Par.Ed_X_ Int. 2		-0.0566		
		(0.156)		
Low Par.Ed_X_ Int. 3		-0.0383		
		(0.161)		
Low Par.Ed		-0.0989		
		(0.112)		
Low Par.Empl_X_ Int. 1			-0.0597	
			(0.154)	
Low Par.Empl_X_ Int. 2			-0.0231	
			(0.179)	
Low Par.Empl_X_ Int. 3			-0.157	
			(0.147)	
Low Par.Empl			-0.0240	
			(0.104)	
Low Par.Empl_X_ Int. 1				-0.223
				(0.154)
Low Par.Empl_X_ Int. 2				-0.218
				(0.173)
Low Par.Empl_X_ Int. 3				-0.201
				(0.157)
Low Par.Empl				0.00827
				(0.112)
Constant	-1.708***	-1.457***	-1.655***	-1.622***
	(0.462)	(0.466)	(0.437)	(0.463)
Observations	3,718	3,725	3,744	3,708
R-squared	0.176	0.177	0.175	0.178

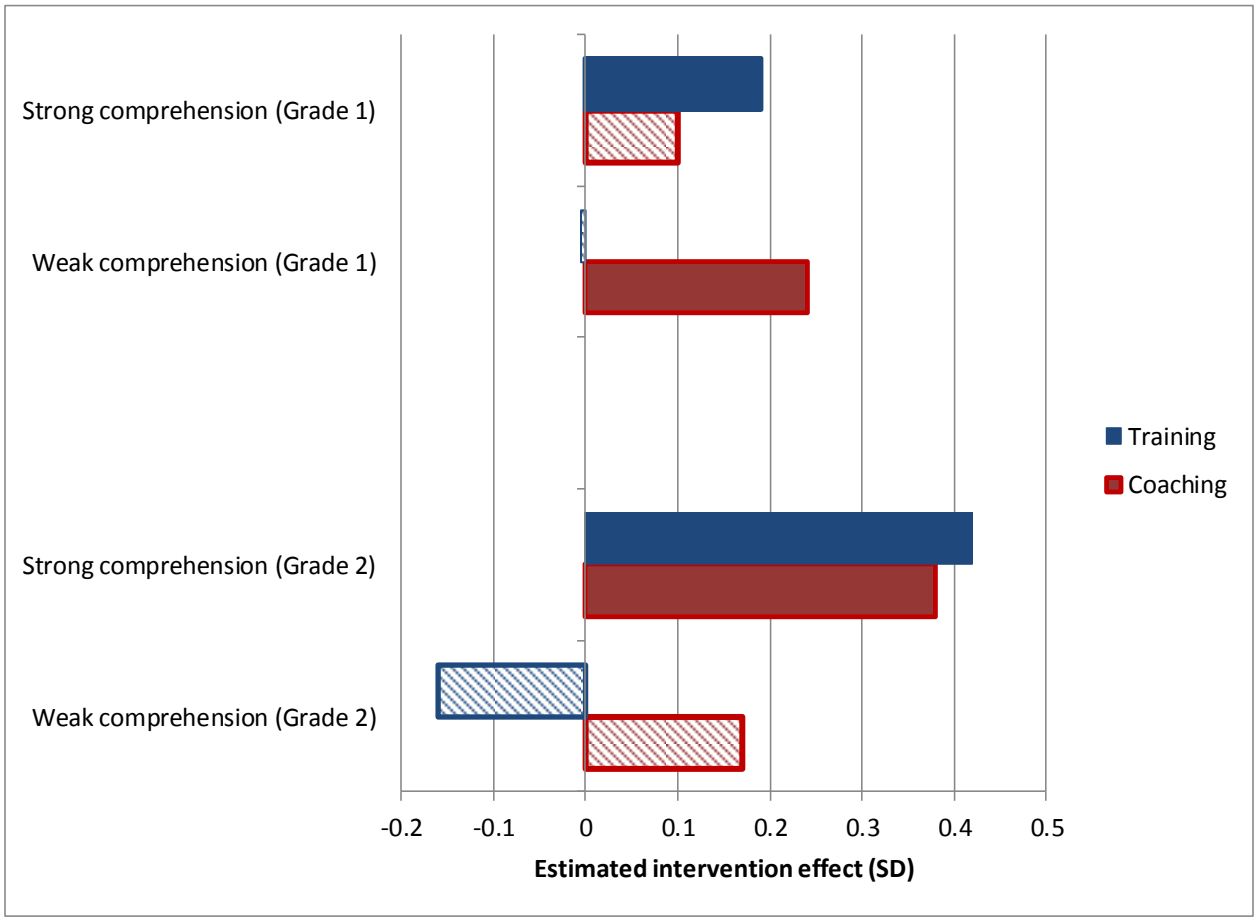
Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

DIFFERENTIAL INTERVENTION EFFECTS BASED ON TEACHER CHARACTERISTICS

We also investigate whether certain teacher characteristics influenced the effectiveness of interventions. Teacher experience and teaching qualifications did not significantly predict the effectiveness of interventions. With respect to teacher absence we observe somewhat conflicting results between the midline (negative interactions between all three treatments and teacher absence) and the endline analysis (positive interactions between two of the treatments and teacher absence) and so overall we're unsure if and how treatment impact depends on teacher absence.

Teacher Setswana reading proficiency was another characteristic hypothesized to potentially influence the effectiveness of interventions. This was measured in two ways. Firstly, we measured the reading fluency of teachers based on self-reported words per minute after reading from a given text in a minute. Secondly, we conducted a short comprehension test. The effectiveness of interventions did not depend on teacher fluency (self-reported words per minute). Figure 34 graphically depicts the estimated effect sizes of the training and coaching interventions for split samples based on a binary distinction in teacher reading comprehension. It is firstly worth noting that teacher comprehension scores were not significantly different across the four treatment groups. We ran two sets of subgroup analysis, first splitting the sample based on Grade 1 teacher comprehension scores and secondly splitting the sample based on grade 2 teacher comprehension scores. In both cases it would appear that the training intervention only worked with teachers who have stronger comprehension scores. However, for the coaching intervention the grade 1 and grade 2 analyses yield somewhat opposing results. In the light of these somewhat inconsistent results as well as the somewhat rudimentary nature of the teacher comprehension test, we feel the evidence around a differential treatment effect based on teacher comprehension scores remains a little thin. Moreover, changes in intermediate outcomes are fairly consistent across the distribution of teacher comprehension scores.

Figure 34: Estimated effects of Interventions 1 and 2 depending on teacher comprehension scores



DIFFERENTIAL INTERVENTION EFFECTS BASED ON PARENT CHARACTERISTICS

Finally we investigated whether any parent or guardian characteristics influenced the effectiveness of the interventions. Table 42 reports the summary statistics for the variables used in this heterogeneous treatment affect analysis. The variable "Mother/Father" is a dummy variable for whether a child's own mother or father completed the parent questionnaire at least once out of the three times it was administered. The "Parent matric" variable is a dummy variable for whether at least one of the child's parents or guardians has completed secondary school education. The variable "Wrote responses" takes a value of one if a child's parent or guardian provided a written response to an open ended question in the parent questionnaire in at least two of the three parent questionnaires that were administered. The "Reads to child" variable and the "Checks homework" variable capture the frequency with which parents reported reading with a child and checking their child's homework, using six categories of responses ranging from "never" to "every day". The "Takes responsibility" variable is a dummy variable

taking a value of one if parents felt that they have the primary responsibility for their child's education as opposed to the teacher, the school or the government.

Table 42: Summary statistics for variables used in parent/guardian subgroup analysis

	count	mean	min	p25	p50	p75	max
Mother/Father	4538	0.74	0	0	1	1	1
Parent Matric	3978	0.31	0	0	0	1	1
Wrote responses	4538	0.45	0	0	0	1	1
Reads to child	3122	4.21	1	3	4	5	6
Checks homework	2712	4.99	1	4	6	6	6
Takes responsibility	2283	0.23	0	0	0	0	1

Table 43 reports the results of regressions investigating whether parental identity, education and parental literacy (as proxied for by writing an open ended response at least twice) had any significant bearing on the treatment effect. Evidently, none of these factors appear to significantly determine the effectiveness of any of the interventions.

Table 43: Intervention effects based on parental identity, education and literacy

	Mother/Father	Parent education	Parent literacy proxy
Intervention 1	0.0643	0.0783	0.0901
	(0.107)	(0.0871)	(0.0929)
Intervention 2	0.197*	0.236***	0.289***
	(0.112)	(0.0899)	(0.0891)
Intervention 3	0.0458	0.0943	0.109
	(0.0968)	(0.0866)	(0.0862)
Mother/Father X Int1	0.0734		
	(0.0882)		
Mother/Father X Int2	0.0658		
	(0.101)		
Mother/Father X Int3	0.0899		
	(0.0917)		
Mother/Father	-0.0463		
	(0.0640)		
Parent Matric X Int1		0.116	
		(0.107)	
Parent Matric X Int2		0.0595	
		(0.101)	
Parent Matric X Int3		0.0700	
		(0.0922)	
Parent Matric		0.0879	
		(0.141)	
Wrote responses X Int1			0.0621
			(0.0989)
Wrote responses X Int2			-0.0757
			(0.0835)
Wrote responses X Int3			0.0279
			(0.0881)
Wrote responses			0.0963*
			(0.0521)
Constant	-1.678***	-1.658***	-1.741***
	(0.453)	(0.476)	(0.455)
Observations	3,781	3,418	3,781
R-squared	0.173	0.161	0.176

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 44 reports the results of regression is investigating whether parental involvement in their child's education (as proxied for by reading with their child, checking homework and taking responsibility for the child's education) influence the effectiveness of interventions. Again, none of these factors appeared to have played a significant role, as evidenced by the non-significant coefficients on the various interaction terms.

Table 44: Intervention effects based on parental involvement in child's education

	Reads to Child	Checks homework	Takes responsibility
Intervention 1	0.217	0.170	0.136
	(0.161)	(0.205)	(0.106)
Intervention 2	0.417***	0.513***	0.207**
	(0.157)	(0.179)	(0.0980)
Intervention 3	0.112	0.286	0.0550
	(0.158)	(0.176)	(0.0978)
Reads to child X Int1	-0.0213		
	(0.0327)		
Reads to child X Int2	-0.0430		
	(0.0340)		
Reads to child X Int3	0.00234		
	(0.0328)		
Reads to child	0.0900***		
	(0.0225)		
Checks homework X Int1		-0.00563	
		(0.0370)	
Checks homework X Int2		-0.0525	
		(0.0334)	
Checks homework X Int3		-0.0272	
		(0.0341)	
Checks homework		0.0959***	
		(0.0208)	
Takes responsibility X Int1			0.0997
			(0.135)
Takes responsibility X Int2			0.0505
			(0.121)
Takes responsibility X Int3			0.0255
			(0.130)
Takes responsibility			-0.105
			(0.0805)
Constant	-1.915***	-1.980***	-1.015*
	(0.488)	(0.522)	(0.526)
Observations	3,059	2,660	1,933
R-squared	0.167	0.168	0.154

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A separate report on family influences on early grade literacy is available and contains a lot more detail regarding the correlates of parent characteristics with early grade learning outcomes, based on the data from the study.

INTERMEDIATE OUTCOMES

Intermediate outcomes for Interventions 1 and 2

In this section we investigate underlying change mechanisms by measuring how the learning environment, teaching practice, and classroom activities changed as a result of the program. For this purpose we draw from three different data-sources: the teacher questionnaire conducted in the full evaluation sample of 230 schools, the classroom and document inspection conducted in the same sample, and detailed lesson observations conducted in a stratified random sub-set of 60 schools.

A preliminary point is important to make: Teachers in the Coaching schools were considerably more likely to report feeling a high level of professional support than those in the control schools, with teachers in the Training group also somewhat more likely to experience high professional support. 82% of teachers in the coaching group felt supported and recognized for their work, compared with 53% of teachers in the control group. Similarly, 84% of teachers in the coaching group reported regularly meeting with people who provide mentoring and curriculum support, compared with 52% of teachers in the control group. These proxies for professional support should to some extent be provided through other teachers in the same school, HODs and subject advisors, but are also a key part of the Theory of Change for the EGRS teacher support interventions, especially the on-site coaching intervention. It is therefore pleasing that as a first step in the process teachers actually felt more support as a result of the interventions.

Table 45: Teachers' experiences of professional support

	Control	Training	Coaching	Parents
I feel supported and recognised for my work	53%	62%	82%	49%
I regularly meet with people who provide mentoring and curriculum support	52%	57%	84%	45%

Two main results are worth emphasizing. First, even though there is no large difference in access to graded readers, the lesson observations reveal that far more pupils are actually reading graded readers in the programme schools. This increase is substantially larger for teachers who received Coaching relative to teachers who received Training (even though they had received the exact same set of reading resources). Second, even though we find no change in the probability *that* pupils practice reading in the classroom, there is a noticeable difference in *how* they practice reading: Teachers in both Training and Coaching arms are more likely to enact group-guided reading, resulting in more opportunities for pupils to receive individual

attention. The impact is, again, larger for teachers who received Coaching relative to Training. These results suggest that there are some teaching practices such as group-guided reading that are difficult to enact and require additional coaching to be effective. They also reveal an important interaction between resources and teaching practice: graded readers are only useful if teachers have developed the skills to use them effectively in the classroom.

We grouped the potential intermediate outcomes into six broad categories: (i) access to reading material in the classroom; (ii) adherence to the teaching routines as prescribed in the curriculum; (iii) curriculum coverage (or the extent of literacy activities conducted) (iv) breadth of reading opportunities in the classroom; (v) teacher-learner interactions related to group-guided reading; and (vi) learners' use of reading material in the classroom. The first two categories – access to reading material and adherence to the teaching routines – provide an indication of at least superficial fidelity to the programme. The third category tests if curriculum coverage has improved because of following the lesson plans. The subsequent two categories look at actual teaching activity in the classroom and tests for the enactment of different components of group-guided reading, an integral yet technically difficult activity prescribed in official curriculum documents. The final category captures what is arguably one of the most important requirements for learning to read: opportunities for pupils to individually practice reading text. For each category we construct a mean index out of the constituent indicators, using the method as specified by Kling and Liebman (2004).

The regression results are reported in Tables 46 to 50. In all specifications we include stratification fixed effects and cluster our standard errors at the school level, where necessary.[11] Many of the variables in this section are ordinal variables that were answered on a 4 or 5-point Likert scale. For ease of interpretation we convert these variables into binary variables when we report the results. In all cases results on statistical significance are essentially the same when running an ordered logistic model on the original variables. The mean index we report is always constructed using the comprehensive ordinal variables, so no information is lost in the mean index.

Access to print and adherence to teaching routine

Row (1) in Table 46 shows that there was a large and statistically significant improvement in overall access to reading material in the classroom: a 0.465 and 0.41 standard deviation increase for the Kling index in the Training and Coaching arms respectively. Rows (2) to (5) show results for indicators that constitute the mean index. There is a substantial increase in the probability that a classroom contains a well-stocked reading corner (a 25 and 26 percentage point increase), and exhibits a sufficient number of quality Setswana posters (25 and 21 percentage point increase) and flash cards on the classroom wall (an 18 and 17 percentage point increase). The magnitude of the impact is remarkably similar for both treatments. It is important to note, though, there is no impact on the probability that every pupil in the classroom

[11] It is not necessary to cluster our standard errors with the subset of 60 teachers where we did lesson observations, because we only observed one teacher per school. The data is therefore already at the school level.

has access to graded readers. This may reflect the complexity for a fieldworker who is not an educational expert to identify a *graded* reader (which refers to a set of booklets progressing incrementally in terms of difficulty) as opposed to any other reading books which may exist in the classroom. The 60-school classroom observation study in fact did reveal significant differences in access to graded readers between control and intervention schools.

Table 46. Access to print and adherence to teaching routine

	VARIABLES	Control	Training		Coaching			p-value Training = Coaches
		mean	Coef.	Std. Error	Coef.	Std. Error	Obs	
	Access to reading material							
(1)	*Kling index*	0	0.465***	(0.120)	0.410***	(0.114)	264	0.651
(2)	All have graded readers	0.416	0.114	(0.0921)	0.0327	(0.0904)	263	0.449
(3)	Reading corners	0.486	0.252***	(0.0854)	0.260***	(0.0806)	253	0.930
(4)	Setswana posters	0.316	0.249***	(0.0821)	0.206**	(0.0865)	263	0.651
(5)	Flash cards	0.752	0.177***	(0.0564)	0.166***	(0.0592)	263	0.828
	Routine							
(6)	*Kling index*	0	0.300***	(0.0811)	0.497***	(0.0652)	276	0.0209
(7)	Group-guided reading	0.241	0.124*	(0.0738)	0.197***	(0.0674)	274	0.363
(8)	Spelling test	0.696	0.155**	(0.0627)	0.238***	(0.0509)	273	0.143
(8)	Phonics	0.491	-0.0708	(0.0745)	0.171**	(0.0720)	274	0.00195
(9)	Shared reading	0.422	0.183**	(0.0728)	0.171**	(0.0711)	274	0.872
(10)	Creative writing	0.310	0.301***	(0.0715)	0.383***	(0.0681)	274	0.286

Notes: Each row represents a separate regression
Cluster robust standard errors in parentheses where necessary
*** p<0.01, ** p<0.05, * p<0.1

Next we test for evidence that teachers are more likely to follow the routine specified in the scripted lesson plans. In the teacher survey, we asked them to report how frequently they conduct various types of teaching activities: group-guided reading, spelling tests, phonics, shared reading, and creative writing.[12] Recall that the frequencies of doing these activities are stipulated in the official curriculum, so in principle the teachers in the control should be performing them at the same frequency. We find that for all of these activities teachers in both Training and Coaching schools are more likely to perform the activity at the *appropriate* frequency. It is important to note that the treated teachers are not stating that they are more likely to perform all activities. They are more likely to perform activities that are required to be performed on a daily basis – group-guided reading and phonics – but state they are less likely than the control group to perform the activity that should only take place on a weekly basis – correcting spelling. These results can therefore not be attributed to pure experimenter demand effect of over-reporting all teaching activities.

[12] Options were: Less than once a week, once a week, 2-4 times a week, every day, twice a day

Group-guided reading

We have learnt that teachers who received the scripted lesson plans appear more likely to follow the right routine, and as a result are more likely to teach phonics and facilitate group-guided reading in the classroom. Next we unpack the type of teaching activities related to group-guided reading. Recall that there are three important components of group-guided reading: individual opportunities to read out loud, individual assessment, and sorting reading groups by ability. We asked about each one of these indicators separately in the teacher questionnaire.

Table 47. Group-guided reading, individual attention, assessment, and sorting by ability

	VARIABLES	Control	Training		Coaching			P-value
		mean	Coef.	Std. error	Coef.	Std. error	Obs	Training = Coaching
	From teacher questionnaire							
(1)	Kling index	0	0.210**	(0.0880)	0.415***	(0.0772)	276	0.0124
(2)	Teacher can provide list of groups	0.430	0.168*	(0.0987)	0.344***	(0.0815)	232	0.0748
(3)	Listens to each pupil read out loud	0.578	0.0324	(0.0772)	0.237***	(0.0638)	273	0.00714
(4)	One-on-one reading assessment	0.655	0.0877	(0.0755)	0.161**	(0.0638)	274	0.296
(5)	Sort groups by ability	0.718	0.107*	(0.0579)	0.144**	(0.0580)	261	0.527
	From lesson observations							
(6)	Kling index	0	0.722***	(0.237)	0.760***	(0.213)	60	0.863
(7)	Pupils split into groups	0.211	0.365**	(0.169)	0.555***	(0.160)	52	0.252
(8)	Pupils read aloud in groups	0.444	0.140	(0.194)	0.410**	(0.158)	54	`
(9)	Pupils read individually to teacher	0.176	0.334*	(0.186)	0.515***	(0.183)	51	0.317
(10)	Individual reading assessment	0.158	0.295*	(0.170)	0.125	(0.177)	55	0.340
(11)	Individual phonics assessment	0.0556	0.175	(0.143)	0.0622	(0.118)	56	0.487
(12)	Reading groups, different texts	0.105	0.0919	(0.133)	0.247	(0.161)	52	0.415

Rows (1) to (5) in Table 47 show results from the teacher questionnaire, which was administered in all 230 schools. There was an overall increase for both treatment arms in the activities that relate to group-guided reading, with a consistently larger impact for Coaching relative to Training. First, as a confirmation of the self-reported increase in conducting group-guided reading, we find that teachers in the Coaching arm are 34.4 and 17.6 percentage points more likely to be able provide a list of the reading groups relative to the Control and Training respectively. We further find that teachers who received Coaching are more likely to state that they listen almost daily to pupils reading out loud (23.7 and 20.4 percentage point increase

compared to Control and Training respectively); more likely to perform one-on-one reading assessment at least weekly (16.1 and 7.3 percentage point increase compared to Control and Training respectively); and more likely to state that they sort groups by ability (14.4 percentage point increase relative to Control). The fact that most of these activities are more likely to take place with teachers who received Coaching vs Training suggests that group-guided reading is a pedagogical skill that requires the additional monitoring and feedback provided from coaches to develop. This is also suggestive evidence that these activities related to group-guided reading are at least part of the explanation for faster acquisition of reading proficiency in the Coaching arm relative to Training.

The results from the teacher survey provide evidence that group-guided reading was more likely to take place in both treatment arms, with the largest increase observed for teachers who received Coaching. Moreover, the larger change seems to come from individual attention, rather than sorting by ability. However, these results are all self-reported. To test if these practices actually changed in the classroom we next turn to results from the lesson observations.

Rows (6) to (12) in Table 47 shows that the results from the teacher survey on group-guided reading are broadly supported by the lesson observations: there is a large, statistically significant increase in the mean index of 0.72 and 0.76 standard deviations in the Training and Coaching groups respectively. When examining the individual indicators that constitute the mean index, we see that there is a large increase in the Coaching arm in the probability that the pupils are split into groups (55.5 percentage point increase), that pupils read aloud in groups (41 percentage point increase), and that the pupils read individually to the teacher (51.5 percentage point increase). The impact for these three indicators is smaller for the Training arm, and not always statistically significant. However, we do not find strong evidence for any improvement in the probability of providing individual assessment and grouping by ability. There is a small increase in the probability of providing individual assessment, which is statistically significant only in the Training arm. Teachers that received Coaching are 24.7 percentage points more likely to have different reading groups assigned to different graded readers (compared to a 9.2 percentage point increase for teachers that received Training), but the difference is not statistically significant.

Taken together we see strong evidence that there was an increase in group-guided reading in both treatment arms, with the largest change observed for teachers that received the Coaching. This coincided with more individual attention by the teacher and opportunities to read out loud in groups, but there is weaker evidence for any change in individual assessment and sorting by ability.

Frequency of opportunities to read

Next, we look at the frequency of reading opportunities in the classroom. The fieldworkers were asked to record how many pupils in the classroom are involved with reading letters, words, sentences, or extended texts. The answers were recorded on a 5-point Likert scale, ranging

from no pupils to all of the pupils. Results are reported in rows (1) to (9) in Table 48. There is only weak evidence that more pupils in the classroom are practicing different reading activities. Pupils in the Training and Coaching arms are more likely to read extended texts, but the mean index is not significantly different across intervention groups. However, it is important to note that these records do not indicate how the pupils are practicing reading. The pupils might be practicing reading through whole-class chorusing, and might not actually have been provided with individual opportunities to read.

Table 48: Opportunities to read

	VARIABLES	Control	Training		Coaching			P-value
		mean	Coef.	Std. error	Coef.	Std. error	Obs	Training = Coaching
	Reading frequency (lesson observations)							
(1)	Kling index	0	0.0767	(0.149)	0.148	(0.150)	60	0.300
(2)	Letters	0.625	-0.126	(0.185)	0.105	(0.174)	49	0.231
(3)	1-2 words	0.471	-0.0408	(0.176)	0.229	(0.227)	44	0.378
(4)	3-10 words	0.667	-0.0582	(0.148)	0.0905	(0.129)	52	0.425
(5)	10+ words	0.133	0.0772	(0.151)	0.111	(0.170)	40	0.406
(6)	1-2 sentences	0.529	-0.269	(0.201)	-0.115	(0.214)	44	0.268
(7)	3-5 sentences	0.333	0.389**	(0.178)	0.441***	(0.161)	48	0.360
(8)	5+ sentences	0.188	0.352**	(0.173)	0.363**	(0.177)	49	0.330
(9)	Extended texts	0.579	0.0262	(0.181)	0.148	(0.182)	55	0.237

Curriculum coverage, assessment and opportunities to write

Table 49 looks at curriculum coverage and teacher assessment of written work. Fieldworkers were required to count the number of days that writing exercises were completed in the exercise book, and the number of pages completed in the government workbook.[13] To minimize risk of bias due to strategic selection of exercise and workbooks, the teacher was asked to provide books of one of the most proficient pupils in his/her class. Table 49 indicates that the amount of written work was higher in both Training and Coaching schools relative to the control group, but there was no statistically significant difference between Training and Coaching schools. The frequency of teachers marking learner work was not significantly different across the groups.

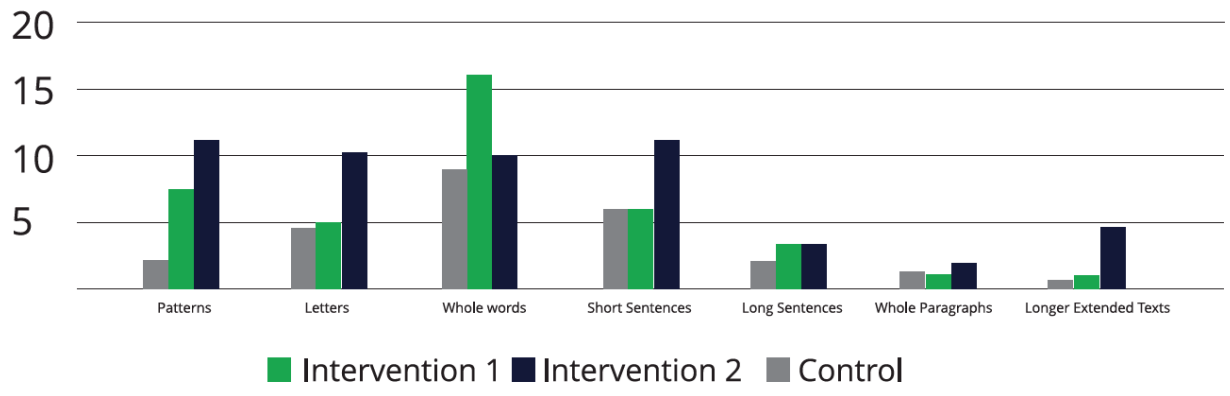
[13] To reduce data capture error, we asked the fieldworker to only count pages completed for three specific days. We chose three days that should have been covered by teachers by the end of the year, regardless of their choice of sequencing.

Table 49: Curriculum coverage and assessments

	VARIABLES	Control	Training		Coaching			P-value Training = Coaching
		mean	Coef.	Std. error	Coef.	Std. error	Obs	
	Curriculum coverage							
(1)	Mean index	0	0.469***	(0.128)	0.317**	(0.139)	271	0.343
(2)	Days pupil completed any exercises	23.57	16.64***	(3.348)	5.007	(3.778)	270	0.00679
(3)	Days pupil completed writing exercises	19.08	8.532***	(3.046)	6.306*	(3.478)	270	0.581
(4)	Days pupil completed full sentence writing exercises	14.11	9.736***	(3.155)	5.539*	(3.044)	270	0.264
(5)	Proportion of pages completed	0.761	-0.0441	(0.0555)	0.0840**	(0.0423)	258	0.0185
	Assessment							
(6)	Learner has marked book	0.939	0.0197	(0.0336)	0.0197	(0.0308)	267	0.999
(7)	All exercises marked	0.400	0.0201	(0.0851)	0.0182	(0.0781)	256	0.984

The 60-school lesson observation data revealed clearer differences across intervention groups with respect to opportunities to write. Figure 35 shows that for most categories of written work there were more writing exercises completed in the exercise books of children in Training and Coaching schools compared to the control group. The more advanced skill of writing extended text was virtually non-existent in control and Training schools, whereas an average of nearly 5 pieces of extended writing was observed in the books of children in the Coaching group.

Figure 35: Average number of writing exercises per writing type



Source: 60-school lesson observation data

Pupil use of reading material

As a final measure, we also test if pupils have more individual opportunities to handle and read books. During the 60-school lesson observation study fieldworkers were required to count how many pupils actually handled books (excluding the government workbooks) and how many pupils read graded reading books during the lesson. Even though there was no difference in access to graded readers between treatment arms, we see a substantial increase in use of reading material, especially in the number of children who have opportunities to read. These results are reported in Table 50. Strikingly, in the control schools only one pupil in one school actually read from a book, leading to an average of 0.05 pupils reading a book in the control. The average number of pupils who read increased by 2.3 and 5.1 in the Training and Coaching arms, respectively. There is also a marked difference between the treatment arms: far more pupils in the Coaching arm handle and read books.

Table 50: Opportunities to handle and read books in class

	VARIABLES	Control	Training		Coaching			P-value
		mean	Coef.	Std. error	Coef.	Std. error	Obs	Training = Coaching
	Use of reading material							
(12)	Kling index	0	4.859*	(2.551)	12.15***	(2.532)	60	0.004
(13)	No. learners handle books	1	0.717	(0.988)	2.542**	(1.001)	59	0.0145
(14)	No. learners read readers	0.0526	2.329**	(1.098)	5.093***	(1.067)	57	0.009

These results reveal the important interaction between resources, teaching practice, and use of resources. Access to graded readers is high in all the evaluation arms, including the control. However the purpose of the graded readers is to provide individual opportunities to practice reading. Pupils are provided this opportunity during group-guided reading, an activity that teachers find challenging to implement. These resources therefore cannot be used without appropriate enactment of a new teaching method. As a result very few pupils are actually reading graded readers in the control schools.

More detail on the intermediate outcomes observed in Training and Coaching groups is available in the separate report on the 60-school classroom observation study.

Intermediate outcomes for parent involvement

The next table investigates the extent to which dimensions of parental involvement may have shifted in response to the intervention, and does so using a set of Ordinary Least Squares (OLS) regressions. Each regression predicts an intermediate outcome that could have shifted. The key explanatory variable is being in Intervention 3 (relative to being in the control group of schools), although a set of control variables for baseline learner scores, learner age and gender, school and community poverty are included (though not reported on in the table). The intermediate outcomes are a parent's frequency of reading to the child, the frequency of checking homework, the frequency of playing games with the child, the number of parent meetings at the school that the parent attended, whether the parent feels that they are primarily responsible for their child's education (as opposed to the school or the government), the frequency of learner absence from school, how often the parent checks their child's school bag, the typical bed time of the child and whether the child sometimes stay up beyond 9pm to watch television.

Table 51 reveals that only the number of parent meetings attended was significantly higher in the Intervention group. This is a somewhat mechanical outcome since regular parent meetings were the mechanism through which any change would have occurred. The fact that no other indicators shifted substantially confirms that on average, there was no large change in parental behavior in response to the intervention.

Table 51: Intermediate outcomes for parent involvement

	(1) readtochild	(2) checkHW	(3) games	(4) Meetings	(5) Responsible	(6) absence	(7) Check bag	(8) Bed time	(9) Late tv
Intervention 3	-0.0182	-0.200	0.0191	0.466***	-0.0364	-0.0766	-0.0284	-0.00381	-0.0299
	(0.119)	(0.140)	(0.0793)	(0.116)	(0.0271)	(0.0674)	(0.0757)	(0.0445)	(0.0307)
Constant	4.217***	4.494***	2.842***	1.423	0.217	2.207***	1.288**	2.011***	0.266
	(0.851)	(1.039)	(0.561)	(0.922)	(0.201)	(0.604)	(0.562)	(0.335)	(0.223)
Observations	1,727	1,502	1,806	1,328	2,574	1,783	1,795	1,427	1,158
R-squared	0.028	0.090	0.020	0.049	0.122	0.038	0.067	0.052	0.049

Notes: Standard set of controls included
Cluster robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

ROBUSTNESS CHECKS

We are not aware of any reasons to expect a systematic upward bias in the estimated treatment effects. Moreover, since attrition is not statistically significantly different across treatment group we do not need to attempt the Lee bounding exercise as proposed in our Pre-Analysis Plan. Therefore, in this section we test the robustness of the main results to two possible issues, both

of which might be expected to attenuate the estimated intervention effect sizes, namely the influence of a few multi-grade classrooms in the sample (where applying the grade-specific lesson plans would be tricky), and the possibility of contamination of control group classrooms due to sharing of lesson plans.

The left hand pane of Figure 36 shows the main results (for those who received two years of interventions – i.e. excluding repeaters) as presented earlier. The right hand pane also excludes the 21 schools (out of 230) where multi-grade teaching occurred according to the teacher questionnaire. As expected, the estimated intervention effects do increase slightly, and most notably for Intervention 1 – we can now be confident of a statistically significant positive impact.

Figure 36: Estimated effect sizes showing 90% and 95% confidence intervals

Note: The standard set of controls is included as in the earlier analysis of main results.

The main risk of contamination of the control group in this experiment arises through the possibility of sharing the EGRS lesson plans. We can rule out the possibility that coaching would have taken place at the control schools or that control school teachers would have attended the central training sessions. Similarly, we know that the additional reading books, flash cards and posters were only provided to treatment schools.

Table 52 shows the percentage of teachers in each of the intervention groups who reported borrowing or sharing lesson plans with teachers in other schools. From the table it is clear that the sharing of lesson plans across schools appears to be a relatively common phenomenon with the percentages ranging between 20 and 30%, although these lesson plans may well include lesson plans other than those provided through the EGRS. This means that control schools may well have, at least to some extent, been exposed to the EGRS lesson plans. Whilst this is only one component of the intervention, it appears important to test the sensitivity of our results to the exclusion of potentially contaminated control schools.

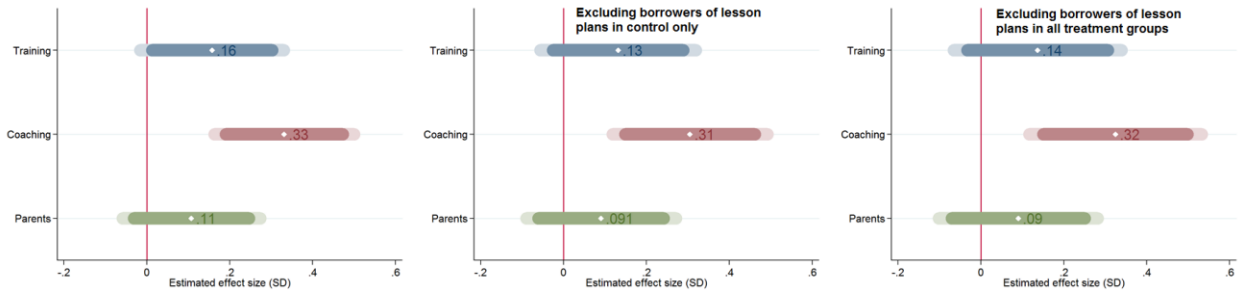
Table 52: Borrowing and sharing of lesson plans across schools

	Receive/borrow lesson plans from other schools	Share lesson plans with other schools
Control	23.28	45.69
Training	19.74	31.58
Coaching	20.73	31.71
Parents	30.38	46.84

Note: The table shows percentages of teachers

Figure 37 diagrammatically presents the results of our main regression models but excluding control schools who reported to have borrowed lesson plans from teachers at other schools, as well as excluding all borrowers of lesson plans in both control and treatment schools. It is clear that the estimated effect sizes hardly change when one excludes borrowers from the analysis. Therefore, there is no evidence that potential contamination of the control group has led to an under-estimation of treatment effects.

Figure 37: Sensitivity of results to exclusion of potentially contaminated control schools



Note: All graphs exclude learners repeating grade 1 since teachers in grade 1 were not interviewed.

12. COST-EFFECTIVENESS ANALYSIS

In thinking about which interventions are suitable to scale up, we need to consider both the impacts and costs of the programmes. In this section we outline different ways of doing cost-effectiveness analysis.

In most scenarios the Parents programme is most cost-effective and Training is least cost-effective. However, since the overall impact of the Parents programme on the full evaluation sample is not statistically distinguishable from zero given conventional levels of significance, we believe it will be irresponsible to recommend scaling up the programme. Moreover Coaching is more cost-effective in producing improvement in the comprehension test, which is arguably the most important goal of literacy. We therefore conclude that the Coaching should be scaled up, since it is the most cost-effective program that is known to work.

For estimates of costs we use the programme budget for the third year of implementation for the Coaching and Training programs. We chose the third year, since this is at a point where a lot of the set-up challenges have been resolved and fixed costs have been paid (all the materials have already been developed). At this point in implementation the largest cost-drivers in both programmes are variable costs (i.e. increase proportionally with number of teachers). One would therefore not expect the difference in per-learner costs to be much different when the programme is scaled up to more schools. Since the programme was not implemented in the third year, we use the budget from the second year of implementation.

The total costs of implementation for the Coaching, Training and Parents interventions were R3M, R2.34M and R0.96M respectively. Since these programmes were each implemented in 50 schools and the average number of grade one pupils in our sample of schools at the start of the programme was 74.6, the per-pupil costs are and R804, R626, and R256 respectively.

Given the impacts of 0.252, 0.12 and 0.1 SD increase for the respective programs, we can conclude that the Parents arm was most cost-effective with a 0.38 SD increase for each R1,000 spent per pupil, compared to 0.31 and 0.18 SD per pupil increase per R1,000 for the Coaching and Training arms respectively.

The above Cost-Effectiveness Analysis looked at the full evaluation sample, but one should arguably use the estimates in the sub-section of schools where the programme can be expected to have the largest impact, namely urban township schools. In these schools the average impacts were 0.76, 0.42 and 0.39 standard deviations in the Coaching, Training and Parents interventions respectively. The Parents intervention remains more cost-effective with 1.5 standard deviations increase per R1000 spent, compared to 0.94 and 0.66 in the Coaching and Training interventions respectively.

Finally we consider another metric of performance, beyond standard deviations increase: the increase in the number of pupils who pass the comprehension test. Now the Coaching is more cost-effective. A learner is 12.3 percentage points more likely to pass the comprehension test per R1000 spent, compared to 6.6 and 3.3 percentage points in the Parents and Training arms respectively.

13. NEXT STEPS IN THE EGRS

Interventions 1 and 2 have continued during 2017. This was done in light of the midline findings emerging during 2016 as we became convinced that it was worthwhile to extend implementation of Interventions 1 and 2 for a third year (grade 3 in 2017) for the following reasons:

1) Based on the midline data from the end of grade 1, both of these pedagogical interventions appeared to be starting to shift reading outcomes, yet the difference between these two interventions was not yet clear.

2. We are learning that system-level interventions such as these, take some time to gain momentum - design aspects of the interventions can be tweaked; relationships with schools need to be developed; coaches become better at their work; etc.

3. The most useful policy evidence would be to present the impact of a full Foundation Phase exposure to these enhanced practice interventions.

We are also planning to conduct a fourth wave of data collection in February 2018. It would have been ideal to do this at the end of 2017 but fundraising and procurement delays have meant that this will not be possible. Beyond this, we plan on using DBE administrative test data and possibly even raising funds for a further round of data collection to measure the longer-term impacts of the interventions.

USING EGRS TOWARDS THE DEVELOPMENT OF READING NORMS IN THE AFRICAN LANGUAGES

The Department of Basic Education, as well as various stakeholders such as the ZENEX Foundation, are interested in developing reading norms for the African languages. Up until this time, detailed work on the development of reading norms in the African languages has not been done in South Africa. Work from elsewhere in the continent is also extremely scarce.

However, the process of developing reading norms is a complex and long-term project. One cannot quickly adapt norms from English, or other languages in which they exist, due to the differing language structures and average word lengths, for instance. One can think of reading norms in the early grades as providing an indication of whether children are on track to reach acceptable levels of learning in later grades. To identify such norms one would therefore really need longitudinal data, that is to say repeated measurements of the same learners, for example from grade 1 to grade 5. One would then define and measure an acceptable level of learning, for example reading comprehension, and then work backwards across the years to see how children needed to be performing in earlier grades in order to be likely to reach the acceptable level of learning in grade 5. This would necessarily amount to an expensive and long-term project.

The department has already invested considerable effort in collecting longitudinal data through the Early Grade Reading Study. We are therefore investigating the possibility of developing reading norms in the African languages by analysing data from the Early Grade Reading Study.

To date, three waves of data collection have already been undertaken in the North West province through the Early Grade Reading Study. At the start of 2015 about 4500 children across 230 schools were assessed in terms of their school readiness at the very beginning of grade 1. At the end of 2015 the same children were again assessed at the end of grade 1. The third wave of data collection was then administered at the end of grade 2 in 2016, yielding three waves of data on the same children. Those children are now in grade 3 and the plan is to administer future data collections when these children are in grade 4 and 5. The focus of the study in the North West has been on home language literacy, which in this case is Setswana. The assessments have been largely based on the well-known Early Grade Reading Assessment (EGRA), but with some adaptations made for the purpose of the study.

Meanwhile a second phase of the Early Grade Reading Study has begun in 2017 in Mpumalanga. The focus of this part of the study has been on English as a First Additional Language, but many items in the assessments are done in the home language since home language literacy is a strong predictor of second language acquisition. Approximately 3500 children were assessed at the start of the grade 1 year in 2017. Donor funds have been raised to assess the same children in November 2017 as well as in November 2018, and the plan is similarly to continue assessing these children up until grade 5 in order to develop a longitudinal dataset. These children will be assessed in English as well as in their home language for the duration of the longitudinal study. The home language for about 70% of the sample is siSwati while the home language for the remainder of the sample is isiZulu.

Analysis of the two Early Grade Reading Study datasets would thus allow for the development of reading norms in Setswana and in siSwati/isiZulu and in that way would cover one language in each of the two main language groups in South Africa.

To our knowledge, the Early Grade Reading Study data is the only suitable dataset for this purpose given its focus on early grade reading in the African languages, it's large sample size and the fact that it is a longitudinal dataset involving tracking the same children over a period of time.

REFERENCES

Adukia, A. (2017). Sanitation and Education. *American Economic Journal: Applied Economics, 9*(2), 23-59.

Altman, E. (2013). *Childhood nutrition in Chile: An assessment of a national school food program.* Atlanta: Thesis submitted to the Emory College of Arts and Sciences of Emory University in partial fulfilment of the requirements of the degree of Bachelor of Arts with Honors.

Andrabi, T., Das, J., & Khwaja, A. (2015). Report cards: The impact of providing schools and child test scores on educational markets. Unpublished.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomised Evaluation in Education in India. *American Economic Journal: Economic Policy, 2*(1), 1-30.

Conn, K. (2017). Identifying effective education interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations. *Review of Educational Research* .

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2011). School inputs, household substitution and test scores. *National Bureau of Economic Research Working Paper 16830*.

Deaton, A., & Cartwright, N. (2016). Understanding and misunderstanding randomised controlled trials. *WP 22595. National Bureau of Economic Research* .

Department of Basic Education. (2012). *Ministerial Audit of Provincial Reading Strategies.*

Diagne, A., Lô, M., Sokhna, O., & Diallo, F. (2014). Evaluation of the impact of school canteen programmes on internal effieciency of schools, cognitive acquisitions and learning capacities of students in rural primary schools in Senegal. . *Partnership for Economic Policy working paper*.

Ebenezer, R., Gunawardena, K., Kumarendran, B., Pathmeswaran, A., Jukes, M., Drake, L., & Silva, N. (2013). Cluster-randomised trial of the impact of school-based deworming and iron supplementation on the cognitive abilities of school children in Sri Lanka's plantation sector. . *Tropical Medicine and International Health, 18*, 942-951.

Fleisch, B., & Schöer, V. (2014). Large-scale instructional reform in the Global South: insights from the mid-point evaluation of the Gauteng Priamry Language and Mathematics Strategy. *South African Journal of Education, 34*(3), 1-12.

Fleisch, B., Schöer, V., Roberts, G., & Thornton, A. (2016). System-wide improvement of early-grade mathematics: New evidence from the Gauteng Primary Language and Mathematics Strategy. *International Journal of Educational Development, 49*, 157-174.

Fleisch, B., Taylor, S., Schöer, V., & Mabogoane, T. (2017). 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. *International Journal of educational Development, 53*, 36-47.

Fleisch, B., Taylor, S., Schöer, V., & Mabogoane, T. (2017). The value of large-scale randomised control trials in system-wide improvement: The case of the Reading Catch-Up Programme. *South African Journal of Education, 37*(1).

Freeman, M., Greene, L., Dreibelbis, R., Saboori, S., Muga, R., Brumback, B., & Rheingans, R. (2011). Assessing the impact of a school-based water treatment, hygiene and sanitation programme on pupil absence in Nyanza Province, Kenya: a cluster-randomized trial. *Tropical Medicine and International Health, 17*(3), 380-391.

Glewwe, P., Hanushek, E., Humpage, S., & Ravina, R. (2014). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In P. Glewwe (Ed.), *Educational Policy in Developing Countries.* Chicago and London: University of Chicago Press.

Glewwe, P., Illias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics, 2*(3), 205-227.

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics, 1*(1), 111-135.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz. (2004). Retrospective vs. prospective analysis of school inputs: the case of flip charts in Kenya. *Journal of Development Economics, 74*, 251-268.

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

Haddaway, N., Woodcock, P., Macura, B., & Collins, A. (2015). Making lierature reviews more reliable through application of lessons from systematic reviews. *Conservation Biology, 29*(6), 1596-1605.

Hanushek, E. (2010). The economic value of higher teacher quality. *National Center for Analysis of Longitudinal Data in Education Research, Working Paper No.56*.

He, F., Linden, L. L., & MacLeod, M. (2009). *A better way to teach children to read? Evidence from a randomized control trial.* Retrieved September 12, 2012, from https://webspace.utexas.edu/linden/www/Teach%20Children%20to%20Read.pdf

Hoadley, U. (2010). *What do we know about teaching and learning in primary schools in South Africa?* Stellenbosch: Appendix B to Van der Berg, S; Meyer, H; Reeves, C; van Wyk, C; Hoadley, U; Bot, M; & Armstrong, P 2010. 'Grade 3 Improvement Project: Main report and Recommendations" for Western Cape Education Department.

Ismail, S., Jarvis, E., & Borja-Vega, C. (2012). *Guyana's hinterland community-based school feeding program (SFP).* Washington, DC: World Bank.

Jukes, M., Zuilkowski, S., Parawan, A., & Lee, S. (2014). Evaluating the cognitive impacts of school health in the Philippines. *International Journal of Child Health and Human Development, 7*, 55-66.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. *Science, 340*, 297-300.

Krishnaratne, S., White, H., & Carpenter, E. (2013). *Quality education for all children? What works in education in developing countries.* New Delhi: International Initiative for Impact Evaluation.

Lieberman, E. P. (2013). *Does Information Lead to More Active Citizenship? Evidence from an Education Intervention in Rural Kenya.* MIT Working Paper.

Luo, R., Shi, Y., Zhang, L., Liu, C., Rozelle, S., Sharbono, B., . . . Martorell, R. (2012). Nutrition and educational performance in rural China's elementary schools: Results of a randomised control trial in Shaanxi province. *Economic Developmnet and Cultural Change, 60*, 735-772.

Machin, S., & McNally, S. (2005). Gender and student achievement in English schools. *Oxford Review of Economic Policy, 21*(3), 357-372.

McEwan, P. (2012). Improving learning in primary schools of developing countries: A meta-analysis of randomised experiments. *Review of Educational Research, 20*(10), 1-42.

McEwan, P. (2013). The impact of Chile's school feeding program on education outcomes. *Economics of Education Review, 32*, 122-139.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 82*, 159-217.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of political economy, 119*(1), 39-77.

Murnane, R., & Ganimian, A. (2014). Improving educational outcomes in developing countries: Lessons from rigorous evaluations. *NBER Working Paper No 20284, National Bureau of Economic Research, Cambridge, MA*.

Nadel, S., & Pritchett, L. (2016). Searching for the Devil in the Details: Learning about Development Program Design. *Worping Paper 434. Center for Global Development*.

National Education Evaluation and Development Unit. (2012). *National Report 2012.*

Nguyen, T., & Lassibille, G. (2008). Improving management in education: Evidence from a randomised experiment in Madagascar. Unpublished manuscript 6.

Nitsaisook, M., & Anderson, L. (1989). An experimental investigation of the effectiveness of inservice teacher education in Thailand. *Teacher and Teacher Education , 5*(4), 287-302.

Nonoyama-Tarumi, Y., & Bredenberg, K. (2009). Impact of school readiness program interventions on children's learning in Cambodia. *Internation Journal of Educational Development, 29*(1), 39-45.

Pandey, P., & Goyal, S. a. (2011). Does information improve school accountability? Results of a large randomised trial. Washington, DC: World Bank.

Pandey, P., Goyal, S., & Sundararaman, V. (2009). Community participation in public schools: Impact of information campaigns in three Indian states. *Education Economics, 17*(3), 355-375.

Piper, B. (2009). *Integrated Education Program: Impact Study of SMRS Using Early Grade Reading Assessment in Three Provinces in South Africa.* Pretoria: United States Agency for International Development.

Piper, B., Zuilkowski, S., & Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Education Development, 37*(11-21), 11-21.

Popova, A., Evans, D., & Arancibia, V. (2016). Inside In-Service Teacher Training: What works and how do we measure it? Washington, DC.: Rise Programme.

Prinsloo, C. (2008). Practice makes perfect - Limpopo study shows the value of exercising literary muscles. *HSRC Review, 6*(4), 5-6.

Ross, K., & Zuze, T. (2004). Traditional and alternative views of school system performance. *IIEP Newsletter*(October-December), 8-9.

Sabarwal, S., Evans, D., & Marshak, A. (2014). The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone. *World Bank Policy Research Working Paper 7021*.

Sailors, M., Hoffman, J., Pearson, P., Beretvas, S., & Matthee, B. (2010). The Effects of First - and Second-Language Instruction in Rural South African Schools. *Bilingual Research Journal*, 21-41.

Simeon, D., Grantham-McGregor, S., Callender, J., & Wong, M. (1995). Treatment of Trichuris trichiura infections improves growth, spelling scores and school attendance in some children. *Community and International Nutrition*, 1875-1883.

Snilstveit, B., Stevenson, J., Menon, R., Philips, D., Gallagher, E., Geleen, M., . . . Jimenez, E. (2016). *The impact of education programmes on learning and school participation in low- and middle-income countries.* London: International Initiative for Impact Evaluation.

Spaull, N., Van der Berg, S., Wills, G., Gustafsson, M., & Kotze, J. (2016). *Laying Firm Foundations: Getting Reading Right.*

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

Taylor, S., & Von Fintel, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. *Economics of Education Review, 50*, 75-89.

Van der Berg, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation, 34*, 145-154.

Watkins, W., Cruz, J., & Pollitt, E. (1996). The effects of deworming on indicators of school performance in Guatemala. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 90*, 156-161.

Wong, H., Shi, Y., Luo, R., Zhang, L., & Rozelle, S. (2014). Improving the health and education of elementary school children in rural China: Iron supplementation versus nutritional training for parents. *Journal of Development Studies, 50*, 502-519.

APPENDIX A: ORIGINAL POWER CALCULATIONS

We will randomly assign 50 schools to each treatment group and 80 schools to the control group, and collect data on 20 grade 1 learners per school. This sample will be sufficient to identify a minimum effect size of 0.21 standard deviations when comparing a treatment group with the control group and a minimum effect size of 0.23 standard deviations when comparing two treatment groups. These calculations assume a 95% confidence interval, an alpha value of 0.8, an intra-class correlation coefficient (rho) of 0.3, and a correlation of 0.7 between pre- and post-test scores. The rho of 0.3 is estimated using the SACMEQ III dataset of pupil reading tests, restricted to the bottom 3 socio-economic quintiles of schools in KwaZulu-Natal. In order to ensure balanced treatment and control groups we will use stratified randomisation. There will be 10 strata, each consisting of 5 schools per treatment group and 8 schools from the control group. Stratification will occur using school performance data from the Annual National Assessments. As Bruhn and McKenzie (2008) demonstrate, stratification using variables that are well correlated to the outcome measure can significantly improve balance and the precision of estimates. Stratification dummies will therefore be included in the regression analysis. Given this procedure to improve precision, the minimum detectable effect sizes referred to above are likely to be conservative.

APPENDIX B: COACH PROFILE

<table>
<tr><th colspan="4">Model Coach Profile</th></tr>
<tr><td>AGE PROFILE</td><td colspan="3">43 – 66</td></tr>
<tr><td>RESOURCES REQUIRED BY EACH COACH</td><td colspan="3">• Car
• Driver’s License
• Laptop
• Cell phone
• E-mail address</td></tr>
<tr><td>APPROXIMATE NUMBER OF SCHOOLS SUPPORTED</td><td colspan="3">17</td></tr>
<tr><td rowspan="2">APPROXIMATE NUMBER OF TEACHERS SUPPORTED PER COACH</td><td>Gr 1</td><td>Gr 2</td><td>Gr 3</td></tr>
<tr><td>32</td><td>29</td><td>28</td></tr>
<tr><td rowspan="2">QUALIFICATIONS REQUIRED</td><td>Coaching</td><td colspan="2">Education</td></tr>
<tr><td>• Coaching and Mentoring Course</td><td colspan="2">At least one of the following:</td></tr>
</table>

	(for example: Matthew Goniwe School of Leadership)	• Undergraduate Certificate (for example: Higher Primary Teaching Certificate; Secondary Education Certificate) • Undergraduate Diploma (for example: Primary Teacher's Diploma) • Undergraduate Degree (for example: BA; BTECH) • Postgraduate Degree (for example: BA Hons; BEd Hons)	
APPROXIMATE NUMBER OF YEARS OF FOUNDATION PHASE TEACHING EXPERIENCE PER GRADE	**Gr 1**	**Gr 2**	**Gr 3**
	16	15	15
APPROXIMATE NUMBER OF YEARS OF FOUNDATION PHASE TEACHING EXPERIENCE PER SUBJECT	**SETSWANA HOME LANGUAGE**	**ENGLISH FIRST ADDITIONAL LANGUAGE**	**MATHEMATICS**
	5	4	4
SKILLS REQUIRED	• Facilitation skills • Computer literacy skills • The ability to develop educational material • Skilled in assessment and moderation • Professionalism • Conflict management skills • Research skills • Organisational skills • Discipline strategies • Time management • Classroom management • Reading strategies		
QUALITIES DEMONSTRATED	• Shows respect for the teachers and their existing teaching skills. • Always has empathy for teachers. • Listens to teachers' opinions about the teaching of reading. • Cooperates with the teachers, school managers and departmental officials. • Challenges teachers' preconceived ideas on the teaching of reading. • Cooperates with teachers, school managers, and departmental officials.		

	• Celebrates teachers' early successes with the programme. • Is always totally truthful with the teachers in terms of their practice. • Creates a safe environment where teachers are able to share their ideas and challenges. • Demonstrates a positive attitude and acts as a positive role model. • Exhibits enthusiasm in the field. • Provides guidance and constructive feedback. • Values the opinions and initiatives of others. • Values ongoing learning and growth in the field. • Expertly demonstrates methodologies and classroom management to teachers. • Sets and meets ongoing personal and professional goals. • Motivates others by setting a good example. • Shares knowledge and experience on the teaching of reading. • Encourages and pressurizes teachers to implement the programme properly.

APPENDIX C: COACHING MANUAL

COACHES MANUAL	
There are a number of coaching practices that are required of each coach, in order to maintain a high standard of professionalism and to ensure that the programme is implemented effectively. This manual intends to guide coaches through these various practices.	
DOSAGE	
DOSAGE REQUIRED: MASTER TRAINING	On average, a coach receives 8 hours of **Master Training** each year. This training is crucial, as it prepares coaches for the **Teacher Training** workshops which coaches provide. The **Master Training** sessions ensure that coaches are thoroughly prepared for these workshops – certifying that each coach has mastered the workshop material, the term's material, the various teaching methodologies and the workshop itself. The **Master Training** is also very important because it focuses on the upcoming term, meticulously preparing coaches for what is expected of them and of the teachers throughout the term. Coaches are briefed on what they are assessing when they walk into the classroom and how they are able to assist teachers throughout the term, ensuring teachers are constantly developing and implementing the programme effectively. Master Training should be provided by high quality expert trainers who are skilled both in **facilitation** and in the **language**

	programme to be used by teachers.
DOSAGE REQUIRED: TEACHER TRAINING	Each coach provides an average of 35 hours of start of term **Teacher Training** and 50 hours of **Afternoon Workshops** each year. **Teacher Training** and **Afternoon Workshops** are important because they help to improve teachers' abilities to teach reading. The training sessions allow teachers to become very familiar with all of the language methodologies. It is during training sessions that teachers are able to demonstrate lessons and get feedback on what they are doing correctly and where there are areas for improvement. The start of term training is also important because it allows for teachers to share their successes with one another – giving teachers the opportunity to take note of good reading practices to implement in their own classrooms. During the afternoon workshops, teachers are given individual attention, which allows coaches to focus on helping teachers to navigate their problem areas as well as further develop their areas of success.

COACHING PROCESS	
OBSERVATIONS	Below is the process that a coach should follow when observing a lesson: *STEP ONE: ORGANISATION ISSUES* 1. At the beginning of each term, it is the coach's responsibility to contact the school and set up a number of appointments, during which, the coach will be observing lessons. *STEP TWO: PROTOCOL ISSUES* 1. When the coach arrives at the school for observations, the first thing that they should do is report to the reception area and sign into the school's administration book. 2. Once the coach has signed in as a guest, they should enquire whether the principal is available. If so, they should greet the principal. 3. Once they have greeted the principal, they should meet the teacher who they are observing to discuss what lesson the teacher is planning on teaching. *STEP THREE: CLASSROOM ISSUES* 1. The coach should observe the teacher's classroom – taking note of whether there is a reading corner, how the classroom is set up, whether there is learning material on the walls and the general condition of the classroom. 2. The coach should then leave the teacher to teach his/her lesson without any interruption. 3. While the teacher is presenting their lesson, the coach should take note of the following areas: - classroom management and discipline - the teacher's preparation level - the teacher's pacing during the lesson - whether the teacher's instructions are clear and coherent - learners engagement throughout the lesson - how efficiently the teacher is sticking to the curriculum, and whether he/she will be able to cover the entire syllabus by the end of the year. - the teachers use of strategies and methodologies set out by the EGRS programme - the positive and negative ways in which the teacher is implementing the programme 4. Before leaving, the coach will monitor the learner's books, ensuring that their written work is at the correct standard and that assessment has been done accurately.
RECORD KEEPING (CLASSROOM OBSERVATIONS)	*STEP FOUR: OBSERVATION RECORDS* 1. The coach should always record the date and what lesson the teacher is presenting. 2. The coach should record what component of the lesson they are observing. 3. The coach should record the positive comments about the lesson – it is crucial that the coach highlights what went well during the lesson. 4. The coach should record two or three areas of improvement. The coach should not overload the teacher with areas of

	concern. 5. The coach should record the teacher's comments and views on the lesson and on the coach's observations. 6. The coach should record his/her suggestions for improvement and next steps for the teacher. 7. The coach should record the teacher's targets for the next lesson. 8. See the example provided
REFLECTION SESSIONS WITH TEACHER	*STEP FIVE: TEACHER REFLECTION* 1. Firstly, the coach will thank the teacher for his her presentation. 2. Before the coach gives his/her comments, he/she will probe the teacher by asking how they felt the lesson went. They will do so by asking questions such as: - How do you feel after the presentation? - What do you think went well during your lesson? - What do you think didn't go so well? - What do you think you could have improved on? 3. After the teacher has given their feedback, the coach will start by telling the teacher what went well in the lesson. The coach well positively narrate the teacher in order to keep them motivated and keep them engaged in the programme. 4. Once the coach has given the positive aspects of the lesson, he/she will go through two or three areas that need development. The coach will not overwhelm the teacher with areas of concern. 5. The coach will give the teacher suggestions for improvement. 6. The coach and the teacher will then set targets for the teacher to achieve by the next lesson observation. 7. The coach will thank the teacher and tell him/her to keep up the hard work and encourage them to keep being persistent.
RECORD KEEPING (TEACHER REFLECTION)	*STEP SIX: REFLECTION RECORDS* 1. Each coach should have a separate notebook for each teacher, or a section in a notebook for each teacher, where they can keep notes and track the teacher's progress methodically. 2. At the end of the year, the coach will easily be able to look at a specific teacher's notes and evaluate the progress that they have made.
COACHING EXAMPLE	**Lesson Observation Notes** **Teacher:** Mrs L Phalatse **Grade:** 3 **School:** Itumeleng Primary School **Day:** Thursday **General Comments**

Mrs Phalatse has kept her classroom clean and well organised. She has a new shelf for the learner workbooks, and she has labelled a space for every subject's books. This is a very good organisational strategy.
At the start of the day, Mrs Phalatse greeted the children, discussed the weather, and let a few children tell their news. She listened attentively when the learners spoke, and she thanked each one of them. It is evident that Mrs Phalatse has a very warm, caring relationship with her learners.
At one stage during the morning routine, there was a disruption when Kgabo, one of the boys, started pushing the other children. I have noticed this kind of behaviour before – this was not a once-off incident. It may be a good idea to add a class rule about not pushing or hitting one another, and to have a discussion with the class about why this is wrong. It is also a good strategy to notice good behaviour, to try and encourage children to follow these actions. For instance, Mrs Phalatse could say: "I notice that Mahlatse is standing still in the line, with his hands at his sides, and his mouth closed. Well done, Mahlatse that is how we enter the classroom."

Phonics
Mrs Phalatse did the correct phonics lesson with learners, for the correct amount of time, 15 minutes. The fact that she is organised and prepared makes the lessons run smoothly, and very little time is wasted. The group monitors collected the exercise books and handed them out, whilst Mrs Phalatse wrote the date and the spelling words on the chalkboard.

Then, Mrs Phalatse revised the phonic sound for the week, and did a quick oral activity to check if learners could spell the words of the week. She made sure she involved many learners in this activity.

Then, Mrs Phalatse asked two children to make sentences with different words. I was pleased to see that the learners used their writing strategies to help them write sentences:

- They said the sentence aloud and counted the words
- They drew a line for each word, and put a full stop at the end
- They wrote the words that they knew
- They used resources to help them write other words
- They said words slowly like a tortoise and wrote down the sounds they could hear.

It is very important to note that the children are able to work independently to write sentences. They used these strategies and did not bother Mrs Phalatse as she listened to Group Guided Reading. Mrs Phalatse has trained her class very well in this aspect – congratulations to her.

I walked around the class and read the sentences that learners were writing, and on the whole, I was very impressed. Most children were able to independently form reasonable sentences. I noticed two children, Mary and Petrus, who really struggled.

- Mary used resources to help her write words, but she could not sound out unknown words. She needs help with phonics, specifically, she needs to work with individual sounds to build up and break down words. Mrs Phalatse must find some time to do regular, individual work with her.

- Petrus is a much more serious case. Petrus battles to hold his pencil and to sit correctly. He also struggled to form letters and clearly could not read or write words. I asked him to sound out a word for me, and he could not do it. This child has serious remedial needs, and should be tested for referral. However, since this could take a very long time, I will help Mrs Phalatse to work out a simple programme of activities that Petrus could do in class. I will suggest the following:
  a. Petrus must sit at a desk on his own, next to Mrs Phalatse's desk at the side of the classroom.
  b. He can participate fully in Listening & Speaking lessons, and in Shared Reading lessons.
  c. For group guided reading, Mrs Phalatse can include him with her small, weakest group, but they must read a very simple text, possibly even Grade 1 level. (She must arrange to get some appropriate Vula Bula books from her Grade 1 colleague.)
  d. Mrs Phalatse must do a separate phonics programme with Petrus – she must get the Grade 1 lesson plans, and must go right back to the beginning, and take Petrus through all the sounds, at a pace that he can cope with. She will have to find time to do these lessons with Petrus whilst the other learners are busy with written work.
  e. During writing lessons, Petrus can work on the same topic, but he can draw a picture of the topic, and then try to write one or two labels, or complete a sentence starter that Mrs Phalatse helps him with.
  f. Mrs Phalatse must try to give Petrus tasks that he can achieve and complete, in order to build his self-confidence.
  g. She must continue to be patient with him and to praise him whenever possible.

Group Guided Reading

Mrs Phalatse has trained her class well in the routine of Group Guided Reading. The orange group knew it was their turn to read, and moved quickly and quietly to the reading area. Mrs Phalatse had the correct Vula Bula books ready for the group.

Mrs Phalatse went through the correct routine for GGR.

- She first told the children what page to turn to.
- Next, she taught them the flashcard words for the day – she did this very nicely, making sure that the learners understood the meaning of all words.
- Then, she asked them to silently read the text on the next three pages, on their own.
- After this, she asked each child to read a portion of the text aloud, on his or her own. She followed carefully, and where a child got stuck, she waited patiently for a bit, before helping them. She helped the children with different word attack skills, rather than just giving them the word. I could see some of the learners implementing the word attack skills on their own – Mrs Phalatse must be commended on this.
- After each child had read, Mrs Phalatse asked the comprehension questions, to check for understanding. She even allowed the group to have an interesting discussion on what they would do in that situation!

I have one suggestion for Mrs Phalatse in terms of GGR – she must update her groups more often, to see that learners are correctly placed in same-ability groups. It was clear that Karabo was a little bored with the pace of this group. Karabo's

	reading has improved nicely, and she clearly comprehends everything that she reads. She is ready to be moved up to the top group in the class. Mrs Phalatse must not let Karabo lose her enthusiasm for reading because she is bored. *Note: I stayed for the full time on Thursday – I watched the next lessons: writing, group guided reading and listening and speaking. Mrs Phalatse is doing a very competent job in all lessons. Her writing lesson was done particularly well. She needs to do some work on her classroom management during the Listening & Spe aking lessons, but I will focus on that during my next observation.* • *For now, I want to congratulate Mrs Phalatse on her excellent work.* • *I also want her to work on the discipline issues mentioned earlier.* • *I want her to set up some strategies to help Mary and Petrus.* • *I also want her to revise her reading groups.* *Thank you Mrs Phalatse, and keep up the excellent work! You are a dedicated, talented teacher and your learners are very lucky to be taught by you!*
AFTERNOON WORKSHOPS	*STEP SEVEN: ADDITIONAL WORKSHOPS* 1. The content that coaches choose to focus on during afternoon workshops is dependent on areas of concern that coaches notice during lesson observations. 2. Afternoon workshops prove to be highly effective because of the individualised attention that teachers receive, the focus on areas of development and the attention the coaches put onto the various strategies and methodologies. 3. Coaches help teachers master the reading strategies and teaching methodologies during these sessions – ensuing that classroom teaching improves drastically.
TRACKING TEACHER DEVELOPMENT	*STEP EIGHT: OVERALL TEACHER DEVELOPMENT* Coaches track teacher's development during the following practices: 1. Training teachers 2. Observing teachers teach 3. Discussing lessons with teachers 4. Facilitating afternoon workshops
ADDITIONAL TEACHER SUPPORT	
PRINCIPALS	It is the coaches' responsibility to keep a good relationship with the principals of the various schools. The coaches should make sure to always greet the principals when they arrive at the school, make sure to have their cooperation, hear their views on the programme, get their views and opinions and ask for advice when appropriate. Coaches should also make sure that teachers are in good standing with the principals. The coaches should ask principals to be involved in classroom matters and to make sure that classrooms are well equipped and resourced. When principals are involved in the classroom, and have a strong presence in the school, the learners are aware that their principal cares about their wellbeing and about their academics. Generally, when the principal is very involved in the school, the school is well functioning and there is a

	good staff relationship. The more involved and supportive the principal of the school is, the more affective the programme will be.
HODS	Coaches invite the HODs to training sessions and to cluster meetings. That way, HODs become familiar with the programme and the methodologies and strategies. When HODs are involved in the programme and are familiar with the programme, they are able to assist teachers when they fall behind or with assessment. The more support teachers have, the more affective teaching becomes.
SUBJECT ADVISORS	Subject advisors are invited to all training sessions. Subject advisors are crucial in assisting in serious matters such as lack of resources, teacher absenteeism, overcrowded classrooms etc. The subject advisors report all such issues to the coaches and coaches are responsible for taking these matters further. Subject advisors are important because they help to keep teachers on task and help them to reach their targeted goals.

APPENDIX D: SUMMARY OF THE 60-SCHOOL LESSON OBSERVATION STUDY

The main strength of Randomised Control Trials (RCTs) is their internal validity in measuring the causal impact of particular programmes. In other words, if outcomes end up higher in a group that received an intervention, we know that this is *because* of the intervention and we can make a quantitative estimate of that impact. But in order to gain a deeper understanding of *why* and *how* a programme may or may not have achieved its desired outcomes, one needs to complement the quantitative estimates of causal impact with mixed methods research. To better understand which mechanisms were affecting the change in Interventions 1 and 2, a classroom observation study was commissioned.

The study was conducted in 60 of the schools that participated in the EGRS. A stratified random sample of 20 schools from each of the Control, Intervention 1 and Intervention 2 groups was chosen to form part of the study. In each of the schools, three different types of evidence were collected: (1) lesson observations; (2) evidence of work done in learners' workbooks and exercise books, as well as the review of various teaching documents and; (3) information from the teacher based on an interview.

Comparing the three different groups of schools, it emerged that the intervention schools were performing notably better than the control schools in the following themes: 'Teaching and Learning Environment'; 'Planning and Curriculum Coverage' and 'Classroom Management'. The main differences in the 'Teaching and Learning Environment' were the increased availability of display material (for example flashcards), a classroom arrangement that is more conducive to reading, and increased availability of reading books in the intervention classrooms.

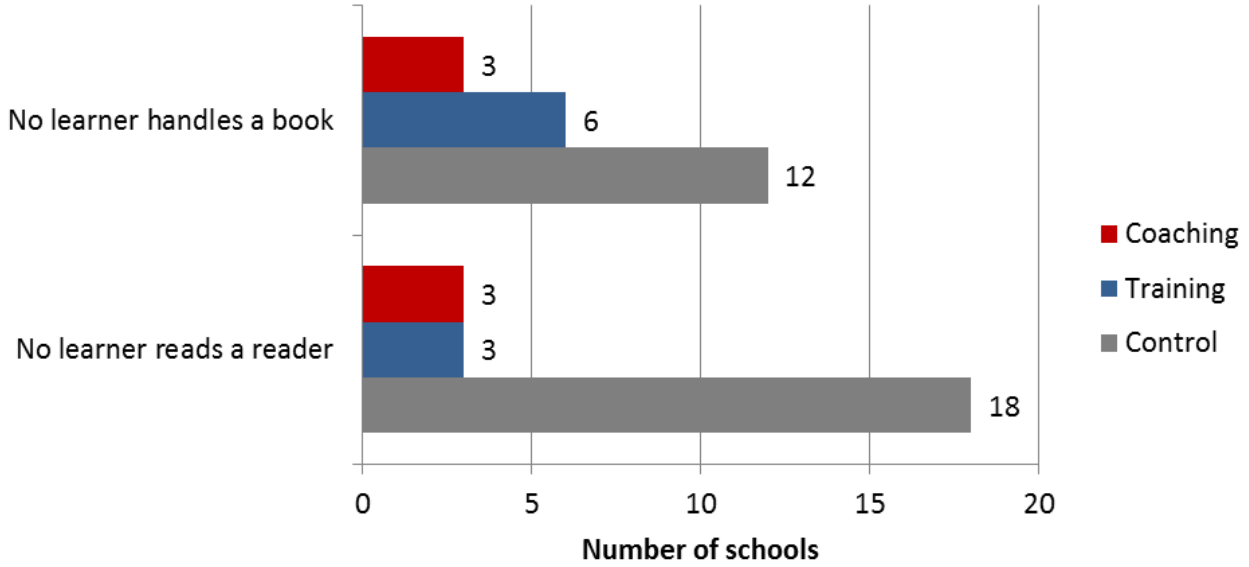
The scripted lesson plans provided through the programme proved to be hugely beneficial in translating the curriculum into daily lessons with detailed activities, which in turn improved 'Planning and Curriculum Coverage'. The specificity of the EGRS lesson plans was visibly different from the lesson plans used by the Control group's teachers and included important aspects such as vocabulary development. The benefit of greater specificity is especially clear with regards to vocabulary development, where teachers in Intervention 1 and 2 schools were much more likely to engage the learners in vocabulary development during the observed lesson. The EGRS lesson plans also provided teachers with a more accurate understanding of the size and scope of the curriculum that needs to be covered across the year, and provided them with a mechanism for tracking their own progress. The teachers in Intervention 2, however, were more likely to actually track their own progress and to be up to date in covering the curriculum. Evidence of increased curriculum coverage in Intervention 2 schools was found in the lessons observed, as well as in the learners' workbooks. The increased curriculum coverage meant that learners were more often engaged in writing activities and therefore learners in the intervention schools were less often observed being uninvolved in class. Although teachers in the intervention schools were observed to have a more realistic understanding of the curriculum

scope, they still did not necessarily have a sufficient understanding of the cognitive demand required by the curriculum.

With regards to classroom management it was found that in 90% of the Intervention 2 classrooms no time was lost due to learners not being involved, whereas this was the case in 75% of the Intervention 1 classrooms. The evidence of more writing exercises in the learners' workbooks in the Intervention 1 and 2 schools corroborates the finding of learners being more involved and suggests that the improved classroom management is leading to increased curriculum coverage. In only 55% of the Control classrooms was no time lost due to learners not being involved.

In relation to the themes 'Opportunities to Write' and 'Use of Learning and Teaching Support Material' there were notable differences between Intervention 1 and Intervention 2 schools. Differences in the 'Use of Learning and Teaching Material' can be largely attributed to the prevalence of learners using storybooks and readers in class, as well as to the use of resources such as flashcards and charts by teachers during lesson observations in the Intervention 2 schools. In 90% of the Control schools not a single learner was observed reading a graded reader, whereas this was commonly observed in the intervention classrooms (see Figure 38). These findings suggest that the EGRS interventions have been successful, not only in providing classrooms with the necessary readers, but also in ensuring that teachers make effective use of these resources.

Figure 38: Use of Learning and Teaching Support Material during lessons



With regards to 'Opportunities to Write', learners in Intervention 2 schools completed more writing exercises on average, specifically exercises pertaining to writing letters, short sentences and extended texts. Learners in Intervention 2 classrooms were also engaged in a wider variety

of writing exercises overall were more likely to have their personally created dictionaries and to do more cursive writing exercises than learners in Intervention 1 classrooms.

Group guided reading provides a valuable opportunity for individualised and small group attention and was observed to occur more often in intervention classrooms. From evidence in the lesson observations it appears that in the majority of Control classrooms, the teachers' in-class reading and phonics assessment was based on the class as a whole, rather than on individual learner proficiency.

The evidence found through the Lesson Observation Study suggests that the reading coaches played a critical role with regard to two aspects: (1) providing teachers with a more in-depth understanding of the enactment of the methodologies they were taught during the training; and perhaps more importantly, (2) supporting and motivating teachers to persist with the implementation of the programme. As mentioned above, there is significant evidence that Intervention 2 teachers were implementing the lesson plans as intended. Intervention 2 teachers were also more frequently seen providing different levels of readers to different ability groups in the lessons observed; doing a wider variety of writing activities during the Home Language lessons; covering the required pages in the DBE workbooks and covering more challenging aspects of the Grade 2 writing curriculum, especially writing sentences and extended texts.

Although Intervention 1 brought about significant changes in teachers' instructional practices, it seems that the reading coach component of Intervention 2 was the essential ingredient to encourage persistence in the curriculum-aligned learning programme. Available evidence therefore suggests that the 'triple cocktail' of lesson plans, high quality materials and coaching is necessary to affect real change in teachers' instructional practices.

APPENDIX E: SUMMARY OF THE CASE STUDIES

A set of case studies was undertaken by Dr Cheryl Reeves in four schools – two Training and two Coaching schools. Each case study involved lesson observations, teacher interviews and document reviews. A full report is available on these case studies.

A number of successful areas of the EGRS programmes were highlighted. Firstly, teachers were making daily use of the EGRS scripted lesson plans and regular use of the EGRS curriculum coverage trackers. Secondly, regular phonics, handwriting, group guided reading instruction and individual seatwork (writing) was taking place in EGRS classrooms. Thirdly, the provisioning of writing activities in the EGRS lesson plans was playing a role in motivating teachers to give classes more writing tasks, and learners were completing written work on most school days.

The case studies also identified several factors inhibiting programme impact. Particularly large classes made it difficult for teachers to provide learners with the individual attention they required. Secondly, there appeared to be an absence of a culture of reading for enjoyment and

limited exposure of grade 2 learners to books besides the graded readers provided through EGRS and the DBE workbooks. Thirdly, teachers displayed a 'restricted' understanding of what it means to teach children to read independently – there was still an over-reliance on teacher-directed strategies (e.g. telling learners what words were).

A second set of case studies was conducted by Dr Kerryn Dixon and Prof Brahm Fleisch in an additional four schools. These were all Coaching schools, selected at the extreme ends of the improvement spectrum based on the average performance on Wave 2 data. As in Dr Reeves' case studies, Dr Dixon and Prof Fleisch observed lessons, interviewed teachers, principals and other school staff, and reviewed classroom documents. A summary report is available on these case studies.

This report focuses on the complexities and nuances associated with the teachers' engagement with the various components and methods of the Coaching intervention. Although teachers lacked the vocabulary to talk about the five components of reading contained in the lesson plans, i.e. Phonological Awareness, Phonics, Vocabulary, Fluency and Comprehension (and writing), the strength of the lesson plans is that they incorporate all of these components in a set of standardised lessons, with simple, systematic routines. The lesson plans impacted both macro (across the academic year) and micro (within each lesson) pacing. Teachers singled out the positive types of learning that occurred during the coaching process, and signaled that a unique and helpful emotional environment was created by the coach. We also found that the new learning materials substantially contributed to improved instruction. The comprehensive set of 'word' flashcards were used extensively. Their popularity may be linked to teachers' familiarity with the 'look and say method' for teaching sight/high frequency words. The Vula Bula books were received very favourably by teachers and were observed in use. Teachers specifically noted that the books were pitched at the correct level and were appropriately sequenced. A number of weaknesses were also observed. The phonics programme was not well understood by teachers. Group-guided reading, a key method for teaching reading was also not properly understood and was inadequately practiced. Whilst group-guided reading was essentially non-existent in Control schools (as evident in the 60-school lesson observation study), this indicates that even in the Coaching intervention there is a long way to go before reaching high quality instructional practice.