Midline I

IMPACT EVALUATION OF FUNDA WANDE COACHING INTERVENTION MIDLINE FINDINGS

MARCH 2020

Prepared by Professor Cally Ardington and Tiaan Meiring

SALDRU
Southern Africa Labour and Development Research Unit

Contents

2 EXECUTIVE SUMMARY 4
2.1. The Intervention 4
2.2. The Evaluation 5
2.3. The Results 6
2.4. Next Steps 8

3 BACKGROUND AND CONTEXT 10

4 PROGRAMME DESCRIPTION 13
4.1 Funda Wande 13
4.2 The Eastern Cape Intervention 14

5 EVALUATION DESIGN 17
5.1 Evaluation Questions 17
5.2 Evaluation Methodology Overview 17
5.3 Recruitment And Random Assignment 18
5.4 Instruments 20
5.5 Ethical Clearance And Departmental Permission 23

6 MIDLINE FIELDWORK 25
6.1 Midline Sample 25
6.2 Attrition And Midline Balance 26

7 MIDLINE RESULTS 28
7.1 Development Of Reading Skills In The Control Group 28
7.2 Treatment Effects 30
7.2.1 Primary Outcome Measures 30
7.2.2 Descripitive Analysis Of Intervention Impacts 34
7.2.3 Regression Analysis Of Intervention Impacts 36
7.2.4 Heterogenous Treatment Effects 46
7.3 Intermediate Outcomes: Teachers 52
7.3.1 Formative Assessment Of Learner Progress In Reading 52
7.3.2 Use Of Vula Bula Anthologies 58

8 DISCUSSION 62

9 NEXT STEPS 72

10 CONCLUSION 74

11 REFERENCES 77

12 APPENDIX 82
12.1. Delivery And Use Of Vula Bula Anthologies In Funda Wande Impact Evaluation Schools 85
12.1.1. Data Sources 85
12.1.2. Delivery And Distribution Of Vula Bula Anthologies 85
12.2. Are Learners Allowed To Take The Vula Bula Anthologies Home? 86
12.3. How Well Used Are The Vula Bula Anthologies? 87
12.4. Are The Vula Bula Anthologies Grade Appropriate? 91
12.5. Future Provision And Training 91

2. Executive Summary

2. Executive Summary

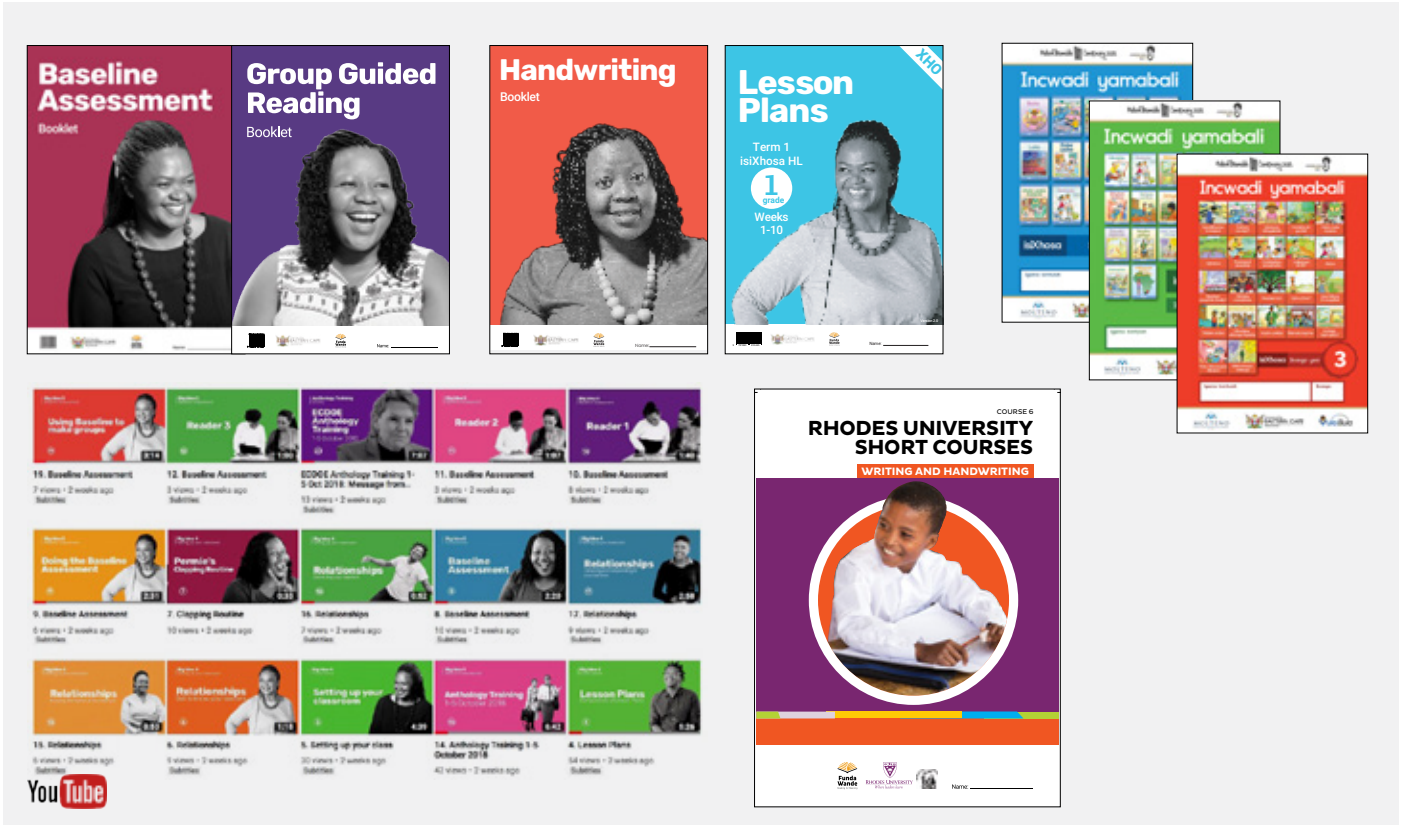
SALDRU at the University of Cape Town is conducting an external impact evaluation of the Funda Wande Coaching Intervention. The primary aim of this evaluation is to assess the causal impact of Funda Wande coaching on foundation phase learners' ability to read with meaning. This research will feed into the growing body of rigorous evidence on which programmes have the potential to address South Africa's reading crisis. This report focuses on the midline findings of the evaluation.

2.1. The Intervention

Acknowledging the limited opportunities for South African teachers to acquire specialized knowledge in teaching reading, particularly in African Languages, Funda Wande has built on lessons from previous interventions and research to design a course to teach reading for meaning in South Africa. The high-quality, multi-media open-access course trains Foundation Phase (Gr R-3) teachers how to teach reading for meaning in African languages. Using professionally filmed in-classroom videos, info-graphics and other multi-media, the course teaches the major components of reading, writing and numeracy in isiXhosa (the pilot language) with subtitles in English *(see **Figure 1**)*. The Funda Wande literacy course and materials have been rigorously developed over two years with input from over 15 South African academics from five universities.

The Funda Wande course is SAQA accredited and has strong support from the national Department of Basic Education, the Eastern Cape

Figure 1: Funda Wande programme components



Department of Education (DBE) and Rhodes University. The "Advanced Certificate in Teaching Foundation Phase Literacy" is a two-year part-time course offered by Rhodes University. In addition to the course, Funda Wande provides on-going coaching for teachers. The Funda Wande in-service training model builds on the DBE's Early Grade Reading Studies' in-service training, which demonstrated the efficacy of using on-site coaching.

Ultimately, the Funda Wande Reading for Meaning project aims to achieve the following goals:

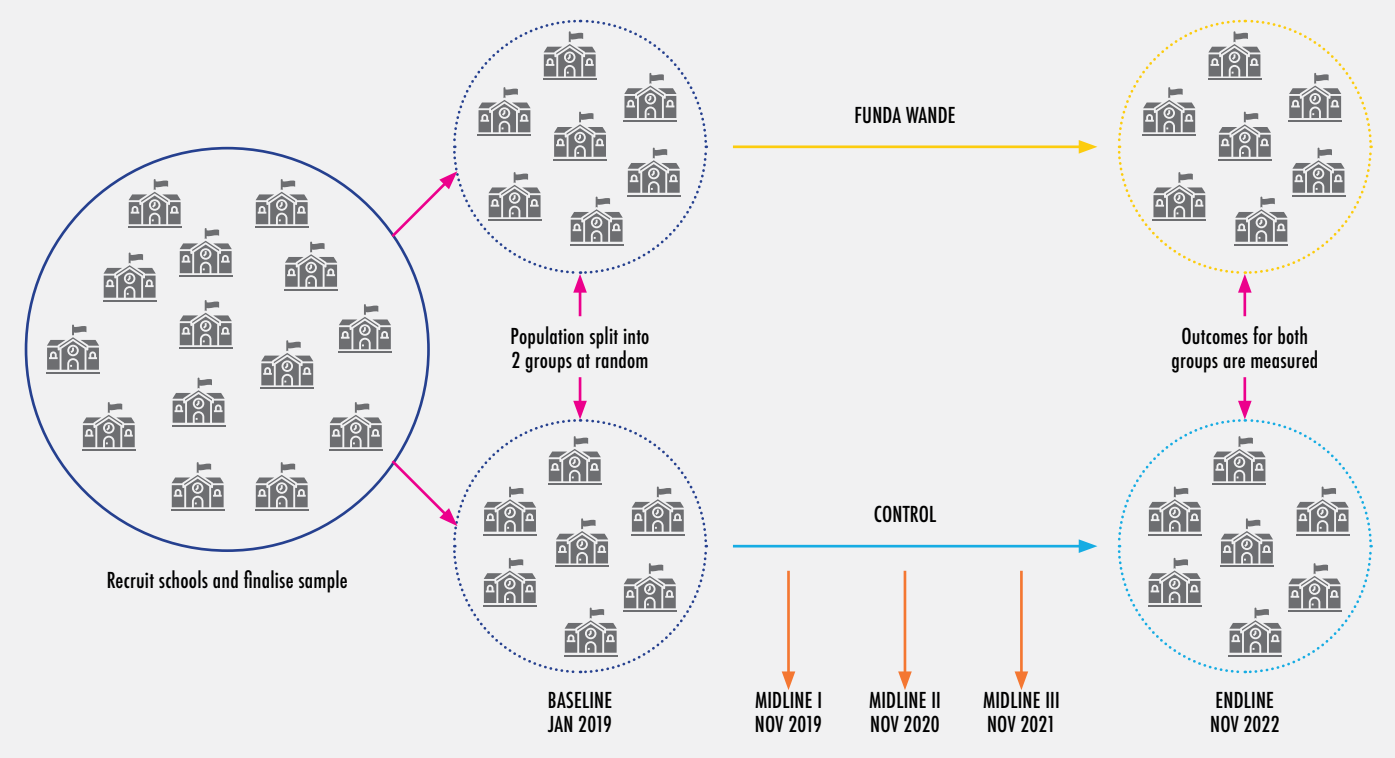
1. All South African children can read for meaning by the end of Grade 3
2. All Grade 1-3 teachers are well equipped to teach children how to read by 2023

2.2. The Evaluation

The impact evaluation uses a randomized control trial (RCT) with schools randomized into one of two arms – Funda Wande and control – for a four-year period (2019-2022). Randomizing the schools into the treatment and control groups ensures that the groups will be very similar before the programme starts. This means that any differences in the reading ability of learners between the two groups at midline and/or endline can be attributed to the Funda Wande programme. ***Figure 2*** provides a graphic representation of the RCT design.

The 29 treatment schools and 30 control schools were selected from urban and peri-urban areas in three education districts in the Eastern Cape: the Nelson Mandela Bay-, Sarah Baartman-, and Buffalo City districts, All schools in the evaluation are no fee, quintile three public schools with an isiXhosa language of learning and teaching. Within each school, 10 learners were randomly selected from both Grade 1 and Grade 2 at baseline. Altogether 1180 of these same learners make up the midline evaluation sample.

Figure 2: Funda Wande Impact Evaluation Design



2.3. The Results

The Funda Wande intervention had a statistically significant 0.17 standard deviation impact on the learner's reading proficiency after the first year. The positive overall effect is driven by an improvement in all the sub-domains of reading proficiency that could reliably be assessed *(see **Figure 3**)*.

In practical terms, learning gains on the subtasks on which the intervention had a positive effect translated to between 20 to 27 percent of a year's worth of learning for Grade 2 learners. At the start of the year, Grade 2 learners scored 31 percent on average for a short reading comprehension assessment. After one year of learning in the business as usual schooling environment, this improved to 46 percent. In Funda Wande schools the statistically significant estimated intervention impact adds another 4 percentage points to Grade 2 learner average comprehension scores – approximately one term's worth of learning *(see **Figure 4**)*

Learners in Grade 1 Funda Wande classrooms gained even more over

Figure 3: Treatment effects for common tasks assessed on both Grades, overall and by sub- task

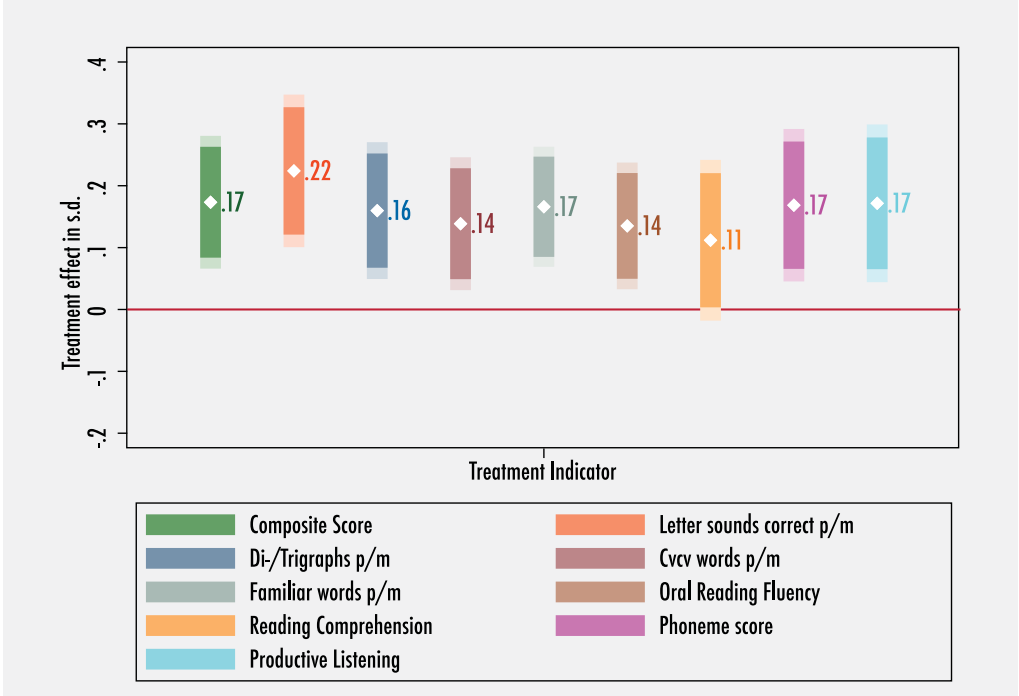


Figure 4: Shifts in the distribution of Grade 2 reading comprehension ability (bin-width= 1 comprehension score)

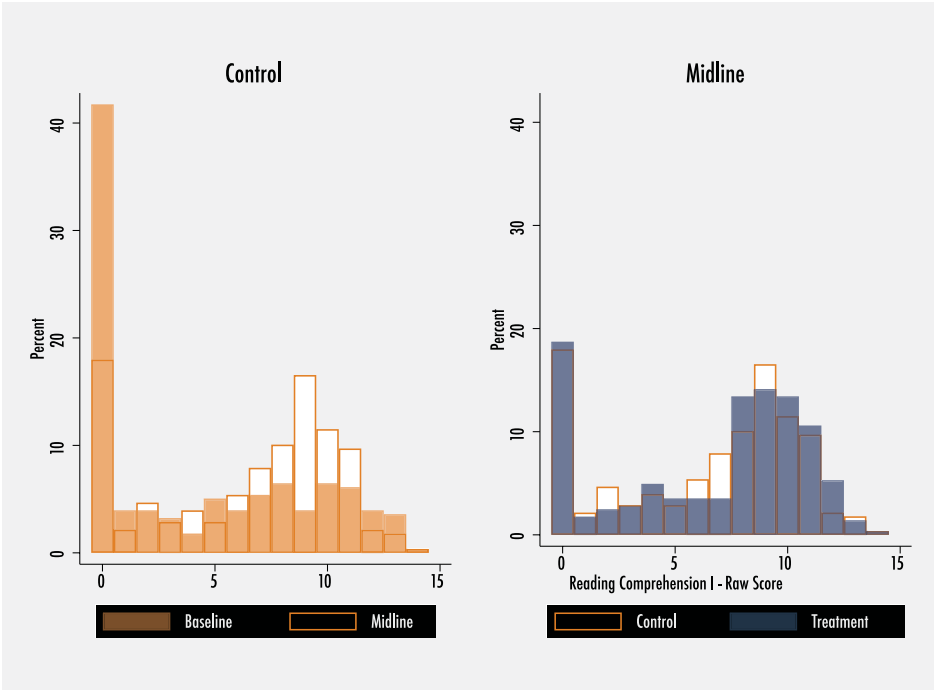
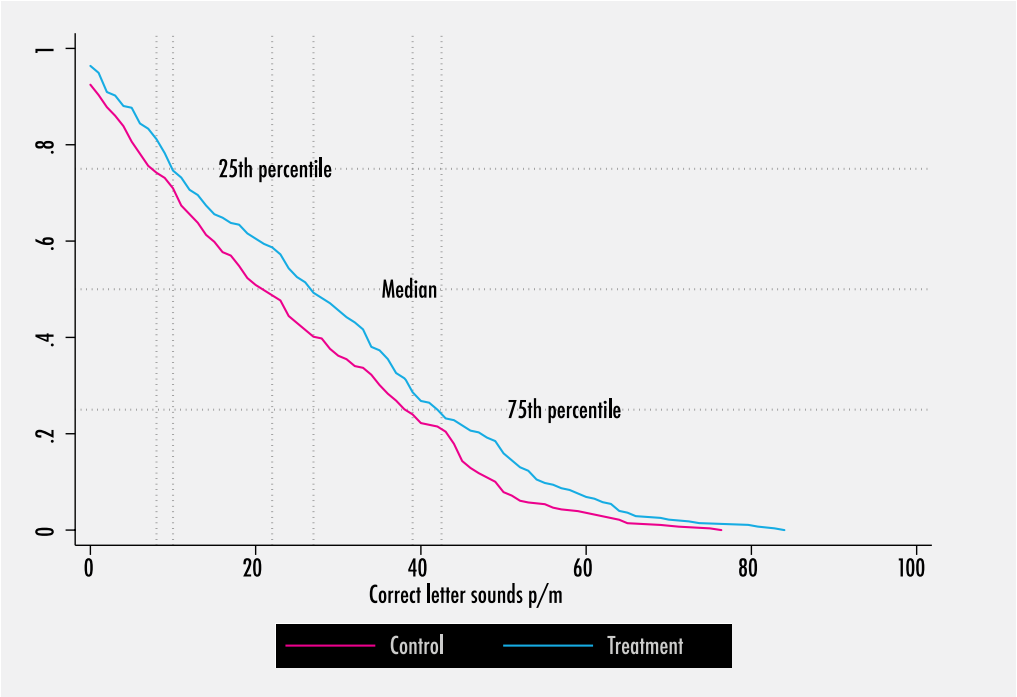


Figure 5: Correct letter sounds per minute by intervention group – Grade 1



their peers in comparison schools for foundational emergent- and pre-literacy skills. For letter recognition tasks, phonemic awareness and productive listening comprehension skills, Grade 1 learners' outcomes improved between 33 to 58 percent of a year of learning more than the status quo gains in comparison schools. Concretely, intervention school Grade 1's could correctly identify on average six letters sounds per minute more after one year of exposure to the intervention, equal to a third of a year's worth of learning.

A particularly encouraging finding from a policy perspective is that the intervention has fairly consistent positive impacts for learners across the distribution of baseline reading proficiency. This is significant, given that previous research suggests that improving reading outcomes for learners with very low foundational reading proficiency levels is particularly challenging (Cilliers et al., 2019). Results for Grade 1 learners letter reading ability is instructive for the programme impacts in general *(see **Figure 5**)*. At the 25th percentile, intervention school Grade 1 learners could read two correct letter sounds per minute more than learners in comparison schools (ten versus eight). At the median there is a five letter sounds difference in favour of intervention school learners (27 versus 22), with a three-and-a-half letter sound difference at the 75th percentiles (42.5 versus 39).

The programme's impacts on Grade 1 learners' foundational skills (letter sound recognition and phonemic

awareness) are particularly large, both relative to the impacts on other Grade 1 literacy skills and the impacts on similar skills for Grade 2 learners. In turn, we only see positive impacts on higher order reading comprehension for learners in Grade 2. Viewed together with other results from the recent literature, these results support the idea that learners require a range of foundational literacy abilities before they can read with some level of speed and accuracy (i.e. fluency), and in turn, they need to read with a certain minimum level of fluency in order to comprehend what they are reading. A potentially related finding is the suggestive evidence that the programme helps boys in Funda Wande schools catch up with their girl classmates, but only in Grade 2 and with the degree of catch-up being contingent on the boys' baseline levels of reading proficiency.

2.4. Next Steps

Looking forward, the positive initial programme impact implies that the next steps include carefully documenting exactly how the programme was implemented in practice, exploring which mechanisms drive the programme effects through qualitative classroom observations, and determining learning gains per rand spent on the Funda Wande programme. These measures will become necessary to understand the cost effectiveness of the programme, as well as its potential to operate at a larger scale. The next midline scheduled for late 2020 will measure the extent to which the early gains reported here compound and continue to translate into improved reading comprehension.

3. Background and Context

3. Background and Context

Despite the substantial progress that has been made towards achieving an almost universal rate of primary school enrolment across the African continent (World Bank, 2018a), levels of actual learning remain low. Education policy researchers, national governments and donors have increasingly focused on how to improve quality of education in developing countries, focusing on what children learn in school and how valuable those skills are once they exit the education system (Piper et al., 2018:1).

Domestically, the 2016 Progress in International Reading Literacy Study (PIRLS) international benchmark test shows that South Africa is unique amongst upper-middle income countries in that 78 percent of Grade 4 children in cannot read for meaning (i.e. at the PIRLS Low International Benchmark[1]) in any language (Howie et al, 2017). South Africa's comparatively poor performance in early Grade reading (and learning outcomes more generally) persist despite an almost universal primary enrolment rate[2], government policies that ensure that the majority of students have access to mother tongue education for the first three years primary schooling[3], and the country's comparatively high expenditure on education by international standards[4] (World Bank, 2018b).

In turn, poor learning outcomes in aggregate are also disproportionately driven by the majority of students who find themselves in the poorest 75 to 90 percent of schools in the country's bimodal education system (Spaull and Pretorius, 2019:155; Spaull, 2013, Fleisch, 2008:21). With large disparities in terms of access to and quality of schooling inputs (both physical and human), functioning accountability structures (like the presence of well-functioning school governing bodies) and eventual learning outcomes, this de facto two tiered schooling system sees children with the largest educational deficits attend schools with disproportionately less capacity (Spaull, 2015; NEEDU, 2013).

The quality of teachers in a child's early years of education seem to have large and persistent effects on both schooling and other later life welfare outcomes, across both developed- (Hanushek and Rikvin, 2010; Chetty et al., 2014) and developing countries (Bau and Das, 2019; Bold, et al., 2017; Bruns and Luque, 2014). However, in South African no-fee public schools[5], teachers generally face the challenge of large and heterogeneous classes, a lack of educational resources/inputs (especially African language readers, textbooks, print, and libraries), and have generally not been sufficiently capacitated with specialized knowledge in teaching reading (due to a lack of initial teacher training or subsequent in-service training) (Van der Berg et al., 2016).

The focus on improving quality has seen an increase in programmes aiming to strengthen teacher capacity through pedagogy focussed teacher professional development programmes in South Africa and elsewhere on the continent [6]. Structured learning programmes have proved successful in bringing about i) instructional change and ii) subsequent improvements in learning outcomes (Snilstveit et al, 2016; Popova et al., 2018). In South Africa, as in

1. If a learner can reach the PIRLS Low International Benchmark then they can "locate and retrieve explicitly stated information, actions or ideas; make straightforward inferences about events and reasons for actions; or begin to interpret story events and central ideas" (Mullis et al, 2017:53).

2. Less than 5 percent of the compulsory age group children are not attending school (World Bank, 2018a:55)

3. More than 70 percent of SA children learn to read in an African language before switching to English in Grade 4 (Pretorius and Spaull, 2016). Thereafter, the majority of South African children (approx. 90 percent) transition to English as language of instruction in Grade 4.

4. South Africa's public expenditure on education is comparable with affluent countries and well above its Sub-Saharan African (SSA) peers, as both a share of total government expenditure, and in per pupil expenditure terms. In 2016/17, the South African government spent about 20 percent of the budget (as a share of consolidated government expenditure) and 7 percent of Gross Domestic Product (GDP) on education. This exceeds both the UNESCO benchmark of 6 percent for developing countries (Motala and Carel, 2019:70), as well as the OECD average of 5.2 percent (IMF, 2019)

5. No fee public schools make up more than 75 percent of schools in South Africa (Spaull, 2019).

Across 33 programs, those programs that link participation to career incentives, have a specific subject focus, incorporate lesson enactment in the training, and include initial face-to-face training tend to show higher student learning gains. In qualitative interviews, program implementers also report follow-up visits as among the most effective characteristics of their professional development program

-Popova et al.

Liberia (Piper and Korda, 2011), Kenya (Piper, 2018) and Uganda (Kerwin and Thornton, 2019), these structured pedagogical interventions have shown early promise. These programmes characteristically consist of (Cilliers et al., 2019, Fleisch et al. 2016):

i) an integrated package approach that include the provision of curriculum aligned learning materials[7] (like Graded readers and other forms of print materials),
ii) teacher guidelines (generally in the form of lesson plans), and
iii) some form of teacher professional development (often comprising of initial teacher training, implementation support, feedback and/or mentoring).

Within the class of structured pedagogy programmes, the mode of delivery has played an important role in programme effectiveness (Popova et al., 2018)[8]. For example, on-site teacher coaching seems to be an especially important component to these programmes' success, both internationally (Kraft et al, 2018) and locally (Cilliers et al., 2019). On the other hand, an important consideration for implementing structured pedagogy programmes at scale are that one-to-one coaching is a very time and resource intensive component of such programmes, and thus a non-negligible obstacle to many resource and capacity constrained governments. Popova et al. (2018) summarise the state of knowledge on the general characteristics of successful teacher professional development programmes:

"Across 33 programs, those programs that link participation to career incentives, have a specific subject focus, incorporate lesson enactment in the training, and include initial face-to-face training tend to show higher student learning gains. In qualitative interviews, program implementers also report follow-up visits as among the most effective characteristics of their professional development program."

In the South African context, current research has built on the lessons learnt and promising insights from previous iterations of similar approaches to improving teacher instruction and student learning outcomes. Examples include the Gauteng Primary Language and Mathematics Strategy (GPLMS) programme (Fleisch et al., 2016 ; Fleisch and Schoër, 2014)[9], the Reading Catch-Up Study (RUCS) Fleisch et al., 2017), the Systematic Method for Reading Success study (Piper, 2009), the Learning for Living project (Sailors et al., 2010)[10] and the Department of Basic Education (DBE)-led Early Grade Reading (EGRS) studies (Cilliers et al. 2019, Kotze et al. 2019).

6. Notably the range of pilot (PRIMR) and at scale (Tusome) studies by Piper, Zuilkowski and colleagues in Kenya (see Piper, 2018; Piper and Zuilkowski, 2014, Piper et al., 2015 and Zuilkowski and Piper, 2017).

7. Often referred to as Learning, Teaching and Studying, or LTSM, materials.

8. As is a recurrent theme in educational interventions, the available evidence suggests that there is more variation in effectiveness across teacher professional development programmes than across classes of educational interventions more broadly (Evans and Popova, 2016; McEwan, 2015).

9. Implemented by the Gauteng Department of Education between 2011 and 2014.

10. All three these studies, however, had specific shortcomings addressed in later EGRS interventions. The former GPLMS did not follow a randomised design, and was post-hoc evaluated by means of quasi-experimental

4. Programme Description

4.1. Funda Wande

Acknowledging the limited opportunities for South African teachers to acquire specialized knowledge in teaching reading, particularly in African Languages, Funda Wande has built on lessons from previous interventions and research to design a course to teach reading for meaning in South Africa. The high-quality, multi-media open-access course trains Foundation Phase (Gr R-3) teachers how to teach reading for meaning in African languages. Using professionally filmed in-classroom videos, info-graphics and other multi-media, the course teaches the major components of reading, writing and numeracy in isiXhosa (the pilot language) with subtitles in English. Through a significant investment by the Allan Gray Orbis Foundation Endowment, FEMEF, Michael & Susan Dell Foundation along with smaller support funding from the VW Community Trust and the Millennium Trust, the Funda Wande literacy course and materials have been rigorously developed over two years with input from over 15 South African academics from five universities.

The Funda Wande course is SAQA accredited and has strong support from the national Department of Basic Education, the Eastern Cape Department of Education and Rhodes University. The *"Advanced Certificate in Teaching Foundation Phase Literacy"* is a two-year part-time course offered by Rhodes University.

In addition to the course, Funda Wande provides on-going coaching for teachers. The Funda Wande in-service training model builds on the DBE's Early Grade Reading Studies' in-service training, which demonstrated the efficacy of using on-site coaching.

Ultimately, the Funda Wande Reading for Meaning project aims to achieve the following goals:

1. All South African children can read for meaning by the end of Grade 3
2. All Grade 1-3 teachers are well

equipped to teach children how to read by 2023

4.2. The Eastern Cape Intervention

Partnering with the Eastern Cape Department of Education (ECDoE), Funda Wande is implementing a pilot intervention to support Foundation Phase literacy teachers in teaching reading for meaning. The essential components of the pilot intervention comprise the following:

Coaching: This comprises of six expert coaches who are experienced foundation-phase literacy educators, resulting in a coach to school ratio of 1:5. The coaches observe Grade 1-3 teachers in their classrooms, provide targeted advice on how to improve their practice, as well as providing model lessons with their students. Coaches visit each school an average of three times a month.

Learner and Teacher Support Material (LTSM) Box: Each teacher is provided with an LTSM box with a set of Funda Wande materials, readers and additional Graded reading aides like posters and phonics flashcards that are aligned to the lesson plans. *(see **Figure 6**)* The Funda Wande materials for teachers include structured lesson plans, handwriting booklets, baseline assessment booklets, group guided reading booklets, online resources for teachers and a pre-loaded flash drive with the full set of Funda Wande videos and multimedia resources. All materials are aligned to the DBE CAPS curriculum and guides. The full-colour Funda Wande lesson plans have one double-page spread per day with photographs of key materials and corresponding guidelines on how to use them.

Training: Training consists of on-site phase meetings once per week and occasional off-site workshops which allow teachers to work collectively on particular issues and to spend time working on Funda Wande materials together, to gain a stronger theoretical understanding of teaching literacy, and to plan for upcoming terms. Training consists of both whole-phase meetings

Figure 6: LTSM Box exhibit

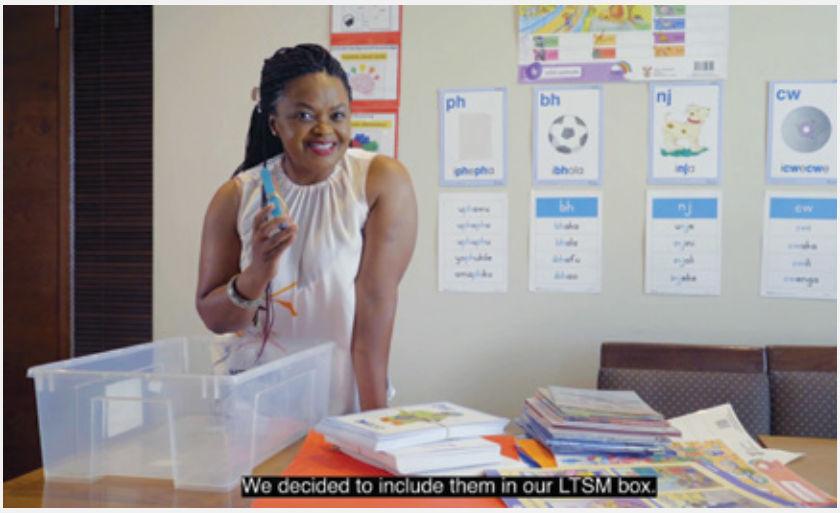


Figure 7: Grade specific lesson plans

Figure 8: Lesson Plan Introduction Section

Figure 9: Lesson Plan Daily Outline Example

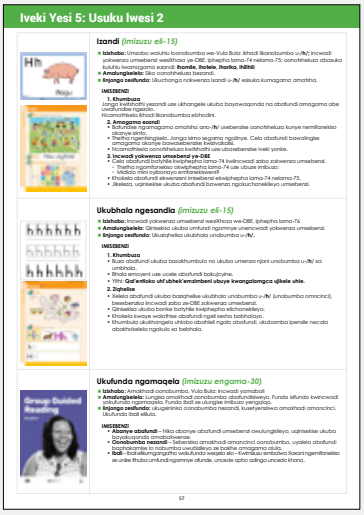


Figure 11: Baseline assessment-, handwriting-, and group guided reading bookletss

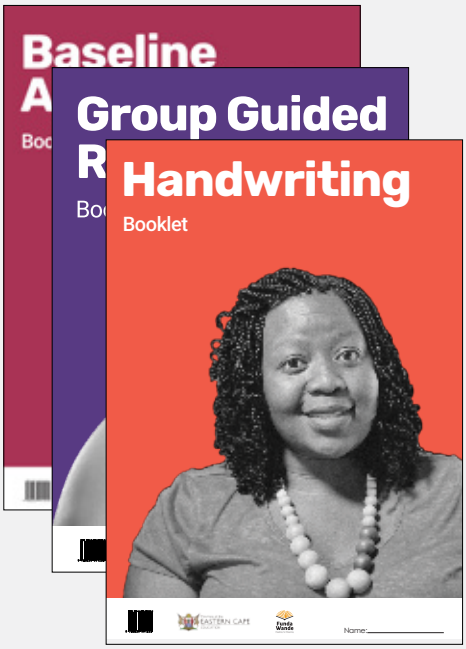


Figure 10: Online Teacher Resources (YouTube training videos)

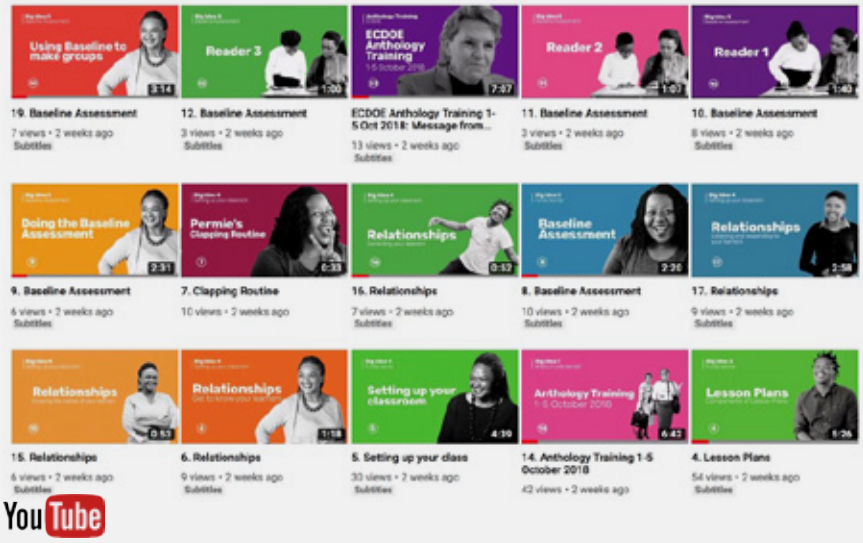


Figure 13: Vula Bula Graded Readers

Figure 12: Group guided reading booklet content example

Figure 14: Training on use of Vula Bula Graded Reader

after school (three per term), and one-on-one in-classroom visits with each teacher in Foundation Phase (at least once per term).

HOD training: It is important that HODs are capacitated to take over the role of coach and literacy specialist after the intervention finishes. To that end all Foundation Phase HODs have been given a bursary by Funda Wande to enrol in the 2-year part-time "Advanced Certificate in Teaching Foundation Phase Literacy" at Rhodes University. This is a blended-learning professional-development qualification that includes block-week sessions at Rhodes as well as off-site work with professional learning communities (PLCs).

The intervention includes all Foundation Phase teachers in each school and was initially scheduled to run for two years (2019-2020). The intervention has recently been extended in both scope and duration to the end of 2022. Section 9 provides details on the expansion of the scope to include support and coaching for numeracy with the same teachers through the *Bala Wande: Calculating with Confidence* programme.

5. Evaluation Design

5.1. Evaluation Questions

The primary aim of the evaluation is to assess whether Funda Wande is effective in moving schools towards the programme's stated goal of all learners reading for meaning by the end of Grade 3. Specifically, we will investigate the impact of the intervention on both foundational reading skills and reading comprehension in the learners' home language.

A secondary aim of the evaluation will be to contribute to ongoing research on how to appropriately measure reading for meaning. This project will also feed into and build on existing empirical research on African languages in South Africa.

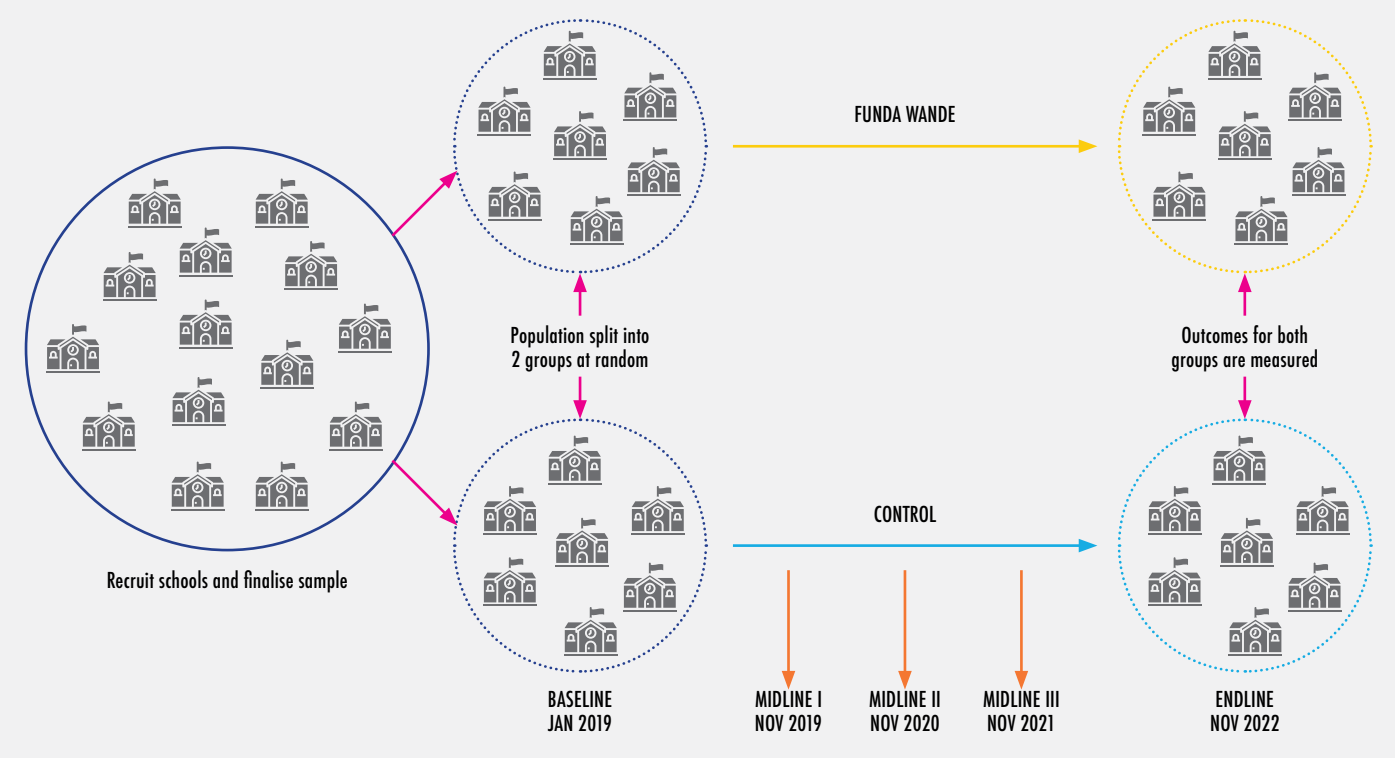
The Eastern Cape pilot will also provide an opportunity for Funda Wande to test and refine the implementation model. Learnings from the evaluation will feed directly into this process.

5.2. Evaluation Methodology Overview

The impact evaluation will use a randomized control trial (RCT) with schools randomized into one of two arms – Funda Wande and control – for a four-year period (2019-2022). The primary hypothesis of the trial is that learners of teachers who receive Funda Wande training materials, resources and coaching support will have better reading outcomes than otherwise comparable learners.

The evaluation team will conduct reading assessments with learners in all schools before the programme starts (baseline) and then four more times over the following four years. We will then compare the average growth in reading skills of learners in Funda Wande schools with learners in the schools that did not receive the intervention. Randomizing the schools into the treatment and control groups ensures that the groups will be very similar before the programme starts. This means that any differences in the reading ability of learners between the two groups at midline and/or endline can be attributed to the Funda Wande programme. The evaluation design and timelines are summarised in ***Figure 15***.

Figure 15: Funda Wande Impact Evaluation Design



5.3. Recruitment And Random Assignment

Working with ECDoE, Funda Wande invited schools from the three urban and peri-urban districts in the Eastern Cape (Nelson Mandela Bay, Sarah Baartman, and Buffalo City) to apply to be part of the programme and then screened the applications to exclude schools with chronic management problems, severe overcrowding (class sizes of 50 plus) or fewer than 20 learners per Grade. These criteria were decided in collaboration with the ECDoE. One approach of the Funda Wande intervention is to "work with the willing", that is to work with schools who do want an intervention in their schools. The logic is that there are thousands of schools requiring support in South Africa and that it makes the most sense to start with schools that want support. All schools are no-fee schools where the vast majority of learners cannot read for meaning. To be eligible a school principal must write a letter of motivation asking to be included in the intervention, with the letter signed by themselves, the Deputy Principal, the HOD and a School Governing Body (SGB) member.

Based on the primary inclusion criteria, Funda Wande received a list of eligible schools from the three respective district managers. The list of 93 schools[11] were based on the explicit criteria that schools should be no-fee, public primary or -combined schools (i.e. have Grade 1-3 learners), with no other major literacy intervention ongoing and an isiXhosa language of learning and teaching (LOLT). Invitations were sent to all 93 schools from the district official lists, of which 77 schools both a) returned completed application forms and b) were self-described as motivated to take part in the study. Funda Wande further screened the applications to exclude schools with chronic management problems, severe overcrowding (greater than 50 learners per class) or fewer than 20 learners per Grade.

Of the returned applications, 63 schools were selected for the programme. From a programme administrative standpoint, Funda Wande also had an informal selection criterion of not including schools that were outside of approximately 1 hour 30 minutes' drive from either of the three central locations (East London, Port Elizabeth or Makana (Grahamstown)).

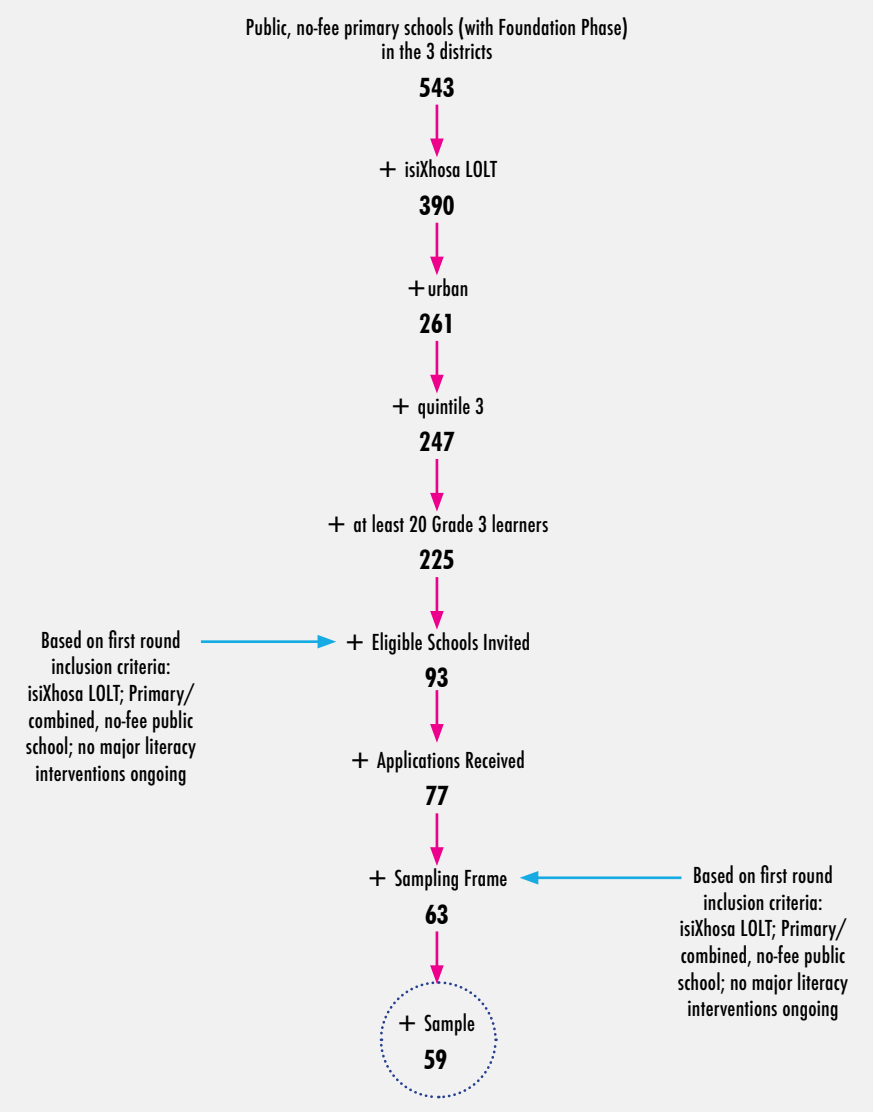


Figure 16: School selection process from plausible populations of schools in the 3 districts

Figure 16 locates the final sample within the universe of public, ordinary, no-fee schools offering Foundation Phase in the three districts. The figures are based on Education Management Information Systems (EMIS) administrative data for the period when the school selection process took place (term three of 2018) merged with data from the Data Driven Districts (DDD) dashboards[12]. Of the 543 schools,

11. Of the 93 schools invited to participate, 36 schools were from Nelson Mandela Bay, 38 from Buffalo City and 19 from the Sarah Baartman district.

12. This database is a collaborative effort by the DBE and the Michael and Susan Dell Foundation, providing education practitioners, -administrators and-researchers with fine-grained learner level data. From this dataset, a school's LOLT was determined to be isiXhosa if all Grade 3 learners had marks for isiXhosa home language (therefore also excluding dual medium schools).

78 percent have isiXhosa as the sole language of learning and teaching (LOLT). Sixty-seven percent of these schools are urban. Almost all (95 percent) of these schools are classified as quintile three. Finally, 91 percent of the remaining schools have at least 20 Grade 3 learners. The total number of schools satisfying these criteria is 225.

The original aim was to have a sample of 10 treatment and 10 control schools per district. The total number of schools in the Sarah Baartman district was only 14. These schools were therefore merged with the Uitenhage schools to create a group of 20 schools. There was a total of 22 schools in the Port Elizabeth district and we randomly selected two of these schools as possible replacement schools, randomly assigning one to treatment and the other to control. Within each group of 20 schools, we randomly assigned half of the schools to receive the Funda Wande program, with the other half serving as control schools. (***Figure 17*** *shows the location of the schools included in the RCT.)*

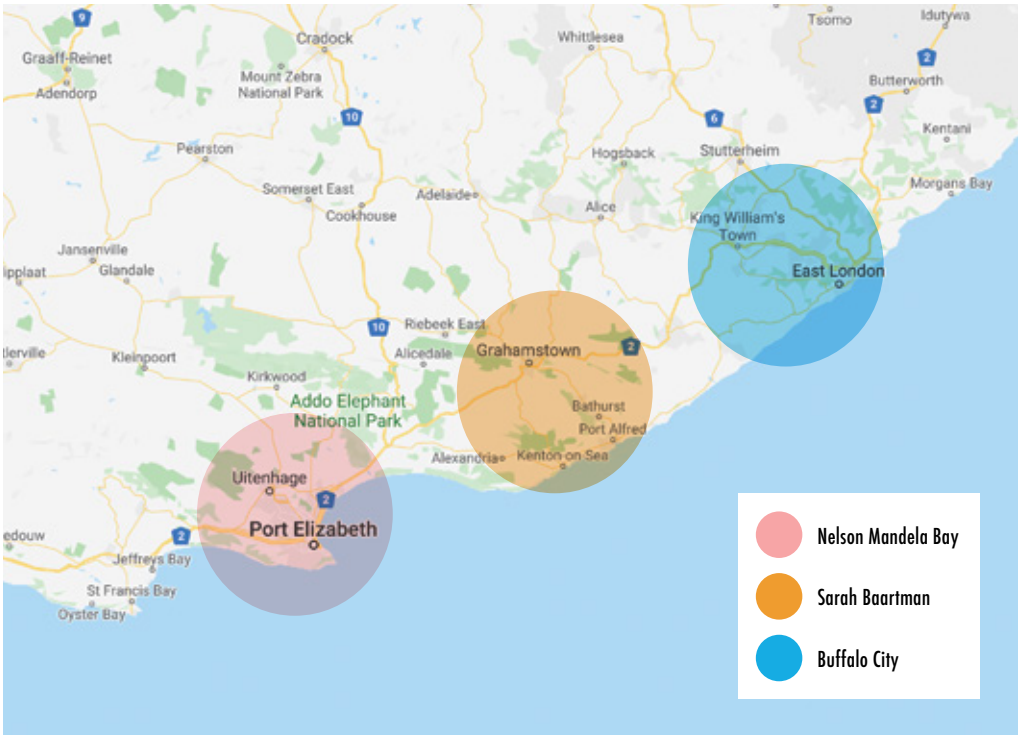
Post randomization, we discovered that the LOLT of two control schools was not isiXhosa throughout the Foundation Phase. We therefore dropped these schools and included the one control school from the replacement group. The final baseline sample is comprised of 29 control schools and 30 treatment schools.

Within each school, we randomly selected one Grade 1 class and one Grade 2 class. Within each of these selected classes, we randomly selected 10 learners. These 1180 learners represent the first two cohorts (A and B). With the extension of the programme, we will add an additional cohort of 10 randomly selected Grade 1 learners in 2020. ***Table 1*** below provides an overview of the three cohorts we will be following over the 2019-2022 period. Although the intervention is in the entire Foundation Phase (Grade R-3), the evaluation is tracking specific cohorts as indicated below:

Table 1: Funda Wande Impact Evaluation cohorts

Grade	2019	2020	2021	2022
1	Cohort A	Cohort C		
2	Cohort B	Cohort A	Cohort C	
3		Cohort B	Cohort A	Cohort C

Figure 17: Location of Funda Wande Impact Evaluation schools



5.4. Instruments

An extended Early Grade Reading Assessment (EGRA) that included a range of pre-literacy and literacy tasks was administered to each of the randomly selected learners at baseline and midline. The assessments used the standard isiXhosa EGRA with adaptations developed by Nangamso Mtsatse, Nwabisa Makaluza and Cally Ardington and drew heavily on the following sources:

- Zenex Foundation (Letter sounds, Phonemic Awareness, Word Reading, Oral Reading Fluency, Reading Comprehension)
- Professor Elizabeth Pretorius (Productive and Receptive Vocabulary)
- EGRS & RTI (Object Naming)
- Room to Read (Sentence Choice)
- ELOM and IDELA (Expressive Vocabulary and Name Writing)
- Wordworks (Writing Letters and Writing Words)
- NORC (Vocabulary)

Many of the tasks built on minor adaptations made by NORC, at the University of Chicago, for the USAID funded Story Powered Schools Impact Evaluation[13]. We have also benefited hugely from the input of the Story Powered Schools and Funda Wande evaluation field teams, particularly on translations and appropriate language.

Table 2 shows the full range of sub-tasks and indicates the Grade(s) and data collection rounds in which they were administered. The inclusion of a range of EGRA subtasks in the baseline assessment was motivated by two key concerns. Firstly, learning to read depends on a complex set of interconnected skills, including both oral language and literacy related skills (Snow, 2017). The evaluation is interested in examining the relationship between these various skills both concurrently and longitudinally as they develop. Understanding where the greatest deficits lie and which skills the intervention most effectively impacts is essential for ongoing programme design. Second, there are statistical reasons for including a range of measures. As the vast majority of learners are not reading at the appropriate level for their age, we expect floor effects (i.e. many learners scoring zero) in many of the core EGRA subtasks, particularly for Grade 1 learners. We therefore employ a range of subtasks, including pre-literacy measures, in order to ensure that there is good discrimination between learners at baseline and at midline.

Understanding where the greatest deficits lie and which skills the intervention most effectively impacts is essential for ongoing programme design.

The range of literacy and pre-literacy assessments conducted at baseline were generally used again for the midline learner evaluations. At baseline, Grade 1 learners were not assessed on certain higher order skills that one would not expect them to have acquired right at the start of their schooling career. However, Grade 1 learners were assessed for most of these skills at midline, including word reading, paragraph reading fluency and reading comprehension tasks. Of the higher order skills from the baseline assessment, only the sentence choice task was conducted on Grade 2 learners only.

A few subtasks from baseline were not included in the midline. The rapid automatized naming (RAN) task was included at baseline to identify learners who had zero, single or double RAN and phonological awareness deficits (Dubek et al. 2017) at baseline, with the interest in tracking the literacy development of these three groups of learners through the waves of the study. The receptive listening task was excluded

13. This is a randomized controlled trial impact evaluation of Nal'ibali's Story Powered School programme involving over 9000 Grade 2 to 4 learners in 360 rural Eastern Cape and KwaZulu-Natal schools. The evaluation runs from early 2017 to late 2019. See Menendez and Ardington (2018).

Table 2: Reading skills and subtasks in Baseline and Midline assessments

Skill	Sub-task & Measurement	Baseline	Midline
Receptive listening comprehension	Perform actions following verbal instruction from the enumerator	Grade 1 & 2	
Productive listening comprehension	Number of questions answered correctly about a passage read aloud by the enumerator	Grade 1 & 2	Grade 1 & 2
Expressive vocabulary	Learner is asked to name items in shop and animals	Grade 1 & 2	Grade 1
Letter sound knowledge	Number of letters sounds identified in 60 seconds	Grade 1 & 2	Grade 1 & 2
Digraph/trigraph sound knowledge	Number of digraphs and trigraphs identified in 60 seconds	Grade 1 & 2	Grade 1 & 2
Phonemic awareness	Identify and manipulate phonemes (starting and ending sounds of words, segmenting words)	Grade 1 & 2	Grade 1 & 2
Word recognition	Selecting the word read by the enumerator from four possible CVCV words	Grade 1	
Rapid Automatized Naming	Number of familiar pictures correctly identified in 60 seconds	Grade 1	
Word recognition	Number of correct CVCV words read in 60 seconds	Grade 2	Grade 1 & 2
Word recognition	Familiar word reading, number of correct words read in 60 seconds	Grade 2	Grade 1 & 2
Oral Reading Fluency	Connected text reading, number of words read correctly from the first reading passage in 60 seconds	Grade 2	Grade 1 & 2
Reading Comprehension	Number of questions answered correctly about the passage read aloud by the student	Grade 2	Grade 1 & 2
Oral Reading Fluency II	Connected text reading, number of words from a second reading passage read correctly in 60 seconds		Grade 2
Reading Comprehension II	Number of questions answered correctly about the passage read aloud by the student		Grade 2
Receptive vocabulary	Identifying correct picture to match word	Grade 1 & 2	
Reading Comprehension	Identifying whether each of 20 short sentences make sense	Grade 2	Grade 2
Writing	Writing name	Grade 1	
Writing	Copying a word	Grade 1	
Writing	Writing letters	Grade 1 & 2	
Writing	Writing words	Grade 2	

from midline due to ceiling effects (i.e. many learners scoring full marks) at baseline. In the interests of avoiding learner fatigue during the assessments, we did not include the baseline writing tasks at midline. Writing tasks are likely to be included in some of the subsequent rounds of data collection.

Grade 2 learners' reading fluency was assessed on two separate passages at midline. The first passage was the same passage used to assess reading fluency and -comprehension at baseline, whilst the newly introduced second passage was slightly longer and more challenging. The first passage was developed by Professor Elizabeth Pretorius for the Zenex Foundation. The newly introduced passage provides a second measure of reading fluency and subsequent reading comprehension assessment for Grade 2 learners, developed by Nangamso Mtsatse and Nwabisa Makaluza from Funda Wande. Both texts are available on request. Having two different

texts on which reading fluency and -comprehension are assessedallows one to go beyond only measuring learners' progression, but also to distinguish whether changes in scores for Grade 2 learners on these tasks are purely down to skills acquired over the academic year (and not to any extent due to learners to recalling the texts). An additional benefit to including the second passage is that it will allow comparisons with SPS sample of rural learners assessed on the same passage (in the first term of Grade 2 and the third term of Grade 3).

For both the passages learners were only asked comprehension questions based on the point up to which they had completed the preceding reading passage. Low levels of reading fluency therefore posed a potential hurdle to assessing reading comprehension: even if learners could at least start reading from the passage, the majority of learners could not read far enough for them to complete all the subsequent comprehension questions (discussed further below). Learners were therefore assessed on their reading fluency based on how many words they could read accurately in the first 60 seconds. Learners were allowed an additional two minutes to continue reading from the two passages, before they were asked comprehension questions based on the respective passages immediately thereafter.

In light of the challenge presented by low reading fluency levels, another task was included to assess learners' reading comprehension whilst relying significantly less on their fluency levels. More precisely, the sentence comprehension subtask was untimed and consisted of 20 short sentences (typically two words in isiXhosa), which learners had to read and then indicate whether the sentence makes sense or not. Each sentence had a pair, for example "Fire is cold" and "Fire is hot". Learners scored one point if their responses for both items in the pair were correct, and scored zero otherwise.

Both the Funda Wande intervention and evaluation therefore place a particular emphasis on the programme's main outcome: reading with comprehension. The measures of comprehension are significantly more extensive than those generally used in Early Grade Reading Assessment (EGRA) type literacy tests (Gove and Wetternberg, 2011). For example, recent adaptations of the EGRA assessments like those in Liberia (Korda and Piper, 2011), Kenya (Piper et al., 2014) and South Africa (Cilliers et al, 2019) all ask four to five short questions subsequent to learners' one reading fluency task. Generally, the sole comprehension task consists of four basic literal questions and one more challenging inferential type question, with learners only asked questions related up to the section of paragraph that they managed to read in the one minute.

Both the Funda Wande intervention and evaluation therefore place a particular emphasis on the programme's main outcome: reading with comprehension.

In contrast, three separate reading comprehension tasks are employed here, all of which provide a more extensive assessment of learners' comprehension skills. The first reading comprehension task is based on a short passage of 41 words. It consists of 14 questions (of which 11 are literal and three are interpretive

questions), with learners are allowed to keep the text in front of them to aid in answering the questions. The second reading passage is longer (55 words in length) and the subsequent comprehension task more challenging. In this case the passage is no longer available as a reference for the comprehension questions and learners are asked ten questions (split halfway between literal and interpretive questions).

At the end of the learner assessment, we conducted a short interview with each learner. At baseline this included questions about books in the home and a range of household possessions. At midline, there were several questions about the Vula Bula anthologies provided by the ECDoE. At each round, the height of each learner was measured and recorded.

At baseline, the field team administered a short questionnaire on access to reading materials and resources with the teachers of the selected Grade 1 and Grade 2 classes in each school. The midline teacher questionnaire focused on the reading proficiency of their class and use of the Vula Bula anthologies. Teachers were also given a task to complete on their own. The task focused on how the teacher would use material from the Vula Bula anthologies in literacy and reading lessons.

At midline, the field team conducted a classroom observation of both teachers' classrooms to determine whether each learner had their own copy of the Vula Bula anthologies. They also recorded whether there were any bookshelves or library corners in the classroom and the extent to which the classroom would be classified as print rich.

Another short questionnaire on other literacy interventions and the delivery of the Vula Bula anthologies was administered to the Principal or HOD at midline.

All assessments and interviews were conducted entirely in isiXhosa by isiXhosa-home-language enumerators.

5.5. Ethical Clearance And Departmental Permission

The study and all associated data collection activities gained approval from the University of Cape Town Commerce Ethics in Research Committee (REC 2018/010/121 and REC 2019/10/007) Permission to conduct research in the schools was also obtained from the ECDoE on 17 January 2019. In addition to the formal application for permission we also created a two page FAQs to inform department officials about the evaluation.

6. Midline Fieldwork

The midline fieldwork was conducted by ikapadata, the same company that conducted the baseline. Despite some challenges, the fieldwork was successfully completed. Full details on training, fieldwork, data cleaning and quality control can be found in the Midline Field Report.

6.1. Midline Sample

Overall, 94 percent of learners assessed at baseline, were re-assessed at midline ***(Table 3)***. The predominant reason for learner attrition were either that learners were absent on the days that fieldworkers visited and revisited the schools, or learners permanently left the school – altogether four percent of the baseline sample. One treatment-assigned school refused to be assessed due to unresolved disputes between the school and Funda Wande. We contacted Funda Wande who confirmed that they had had a very difficult experience with this school and that their coach was unable to visit the school due to objections about scheduling. The ECDoE also confirmed that this was a "problem school.". After the multiple attempts to engage the school, we had no choice but to remove this school. This accounts for a further two percent of the baseline sample. In the other isolated cases (less than one percent of the baseline learners), learners were not reassessed because they refused reassessment, they had behavioural and/or learning disabilities prevented assessments from taking place, or the assessment was not captured due to technical challenges in data collection[14].

Hearing and vision screening was successfully completed for 87 percent of Grade 1 and 90 percent of Grade 2 learners who completed the EGRA assessments and interviews.

Table 3: Midline status for full sample of learners

	Grade 1	Grade 2	Total
Assessed	555	561	1116
School refused	10	10	20
Learner no longer at school	14	7	21
Learner absent	12	11	23
Learner refused	2	1	3
Special needs	2	0	2
Data error	0	2	2
Total	595	592	1187

Of the 118 teachers interviewed at baseline, 19 could not be interviewed at midline. Substitute teachers replaced baseline teachers if the baseline teachers met any of the following replacement conditions: a) they moved from the school/retired, b) they are teaching a different Grade in the school, or c) if they were on sick/incapacity leave starting prior to commencement of fieldwork and with unknown return date. These teachers were replaced with the current class teacher for the class selected at baseline. Nine teachers were replaced. The remaining seven teachers were in the school that refused the field team access (two teachers), could not be interviewed due to time constraints (four teachers), or were absent on all visits from the field team (four teachers).

Interviews were conducted with principals or HODs in 57 schools. In

14. The original sample design and power calculations were based on assessing 10 Grade 1 and 10 Grade 2 learners at each school. At the start of baseline fieldwork, we decided to explore whether it would be possible to complete 12 learners in each Grade. Within the first week, it became clear that this was an unrealistic target and we reverted back to the original plan of 10 learners per Grade. Baseline data includes 11 schools with 12 Grade 1 learners, 2 schools with 12 Grade 2 learners and 3 schools with 11 Grade 2 learners. For these schools, the additional learners assessed at baseline were used as replacements for unavailable (absent/transferred/refused) learners. Table 3 includes seven of the additional learners assessed at baseline who were used as replacement learners. These seven learners replaced four absent learners and three learners who had left the school.

one school, neither the principal nor the HOD was available for interview during either the first fieldwork visit or the mop-up visit. The interview is also missing for the school that refused to participate altogether. Classroom observational checks on the Vula Bula anthologies were conducted in 55 Grade 1 and 55 Grade 2 classrooms. Aside from the school that refused to participate, classroom observations were not conducted due to non-availability of teachers and/or learners.

Overall, 94 percent of learners assessed at baseline, were re-assessed at midline.

6.2. Attrition And Midline Balance

The overall attrition rate for the learner sample was six percent[15]. Attrition was slightly higher amongst learners in treatment schools as compared to learners in controls schools (seven percent versus five percent).

Attrition has two potential impacts on the RCT. The first is a small reduction in statistical power with a slightly smaller sample. Our power calculations behind the sample design are based on fairly conservative assumptions so this reduction in sample size is not of great concern. The second is the potential for selection bias to be introduced into the sample, thereby threatening a key strength of the RCT methodology: the internal validity of the estimated programme impacts.

Baseline equivalence of the full sample was guaranteed through the random assignment of schools to treatment and control using statistical software. Subsequent checks to assess attrition suggest that any differential attrition between learners in treatment and control schools does not lead to an imbalanced sample at midline. Column (1) in appendix ***Table A1*** shows the regression of treatment status on whether or not a learner attrited, taking into account the sub-district (or strata) within which schools were randomly assigned. There is no statistically significant differential attrition between learners in the two groups. The low overall attrition and minimal differential attrition[16] comfortably fall within the What Works Clearinghouse (WWC) (2020:10) conservative limits for low expected bias, and therefore meets the highest possible WWC standards.

In a similar vein, appendix ***Table A2*** shows that non-attriting learners are still balanced in terms of a range of baseline assessment scores, observable learner characteristics and household assets. Standard protocols are available to assess whether the two groups are equivalent at midline, using the standardised mean differences (of effect sizes[17]) in variables between treatment and control groups (WWC, 2020:13). As expected, the differences between treatment and control on all variables are within the limits to satisfy baseline equivalence[18]. Given the checks on the balance and equivalence of the two groups of non-attriting learners, we are satisfied that the programme impacts can be reliably estimated.

As is to be expected from previous evaluations (for e.g. Cilliers et al., 2019), teacher attrition was slightly higher at eight percent[19]. Although the attrition rate was higher in treatment schools than control schools (10 percent versus seven percent), this difference is not statistically significant.

15. The attrition rate takes into account the seven learners who were replaced (i.e. attrition is calculated for the sample of 1187 learners).

16. See Table A1, column (1): there is a three percent mean difference in attrition rates between learners in the two groups.

17. Effect sizes are calculated as the difference in means between the treatment and control groups, divided by the pooled standard deviation for the variable.

18. If the effect sizes are 0.05 or less in absolute value, the two groups are considered equivalent on that dimension. When effect sizes are in the range between 0.05 and 0.25, the baseline measures should be included as controls in the model estimating programme effects to satisfy equivalence Variables for which such adjustments are required include 12 of the 18 sub-tasks, learner age, whether they have non-academic books at home, and whether their household owns a computer or has a toilet. These variables are therefore added as controls to satisfy equivalence between the two groups, and not only to improve the precision of the estimates of programme impact. No effect size is greater than 0.25 in absolute value - the level at which the samples are considered to be not equivalent.

19. This proportion includes substitute teachers as non-attritors as the information that we collect from teachers is about the availability and use of reading materials in the class rather than information about the individual teacher. Excluding the replacement teachers, the attrition rate is 16%.

7. Midline Results

7. Midline Results

The range of subtests provide insights into both i) the literacy and pre-literacy sub-tasks that learners can and can't do, ii) how these skills develop over time in the status quo schooling environment, and iii) how these skills are affected by the Funda Wande intervention. We begin by examining the development of these skills amongst learners in the control schools before moving on to quantifying the impact of the Funda Wande intervention.

7.1. Development Of Reading Skills In The Control Group

Table 4 presents a summary of the average midline score, standard deviation and percentage of learners scoring zero on each of the EGRA sub-tasks for Grade 1 and Grade 2 learners in control schools. For all the tasks that were conducted at both baseline and endline, we also summarize the change in the percentage of control group learners scoring zero and the average scores in ***Figures 18 and 19.***

The percentage of Grade 1's scoring zero for letter sounds decreased from more than half of the sample at the start of the year, to only six percent by year end. For the more challenging letter sounds (digraphs and trigraphs), 58 percent of Grade 1's could not sound out one digraph or trigraph by the end of the year (down from 97 percent). Reading digraphs and trigraphs are a crucial foundational skill for reading words in isiXhosa. At the other end of the spectrum, about one third of Grade 1's could correctly sound more than 33 sounds per minute correctly, falling to only one in twenty for digraphs and trigraphs.

Table 4: Midline EGRA score for control group learners

	Grade 1			Grade 2		
	Mean	Std dev.	% zero score	Mean	Std dev.	% zero score
Correct letter sounds per minute	24.3	18.5	8%	44.8	20.3	1%
Correct di/tri-graphs per minute	6.2	10.6	58%	24.3	19.6	15%
Phonemic awareness	3.3	2.1	10%	5.0	1.9	2%
Productive listening comprehension	3.2	1.4	4%	3.9	1.2	1%
Correct CVCV words per minute	6.3	9.5	50%	20.4	16.4	20%
Correct words per minute	4.2	6.7	50%	14.7	11.9	22%
Oral reading fluency (passage 1)	4.5	7.6	50%	16.7	14.0	16%
Reading comprehension (passage 1)	2.3	3.3	58%	6.4	4.0	18%
Expressive vocabulary	9.1	3.5	0%			
Vocabulary				5.5	0.8	0%
Sentence choice				5.4	3.8	27%
Oral reading fluency (passage 2)				15.3	13.5	24%
Reading comprehension (passage 2)				4.0	3.1	27%
Observations	279			278		

Figure 18. Percentage of control group learners scoring zero at baseline and midline

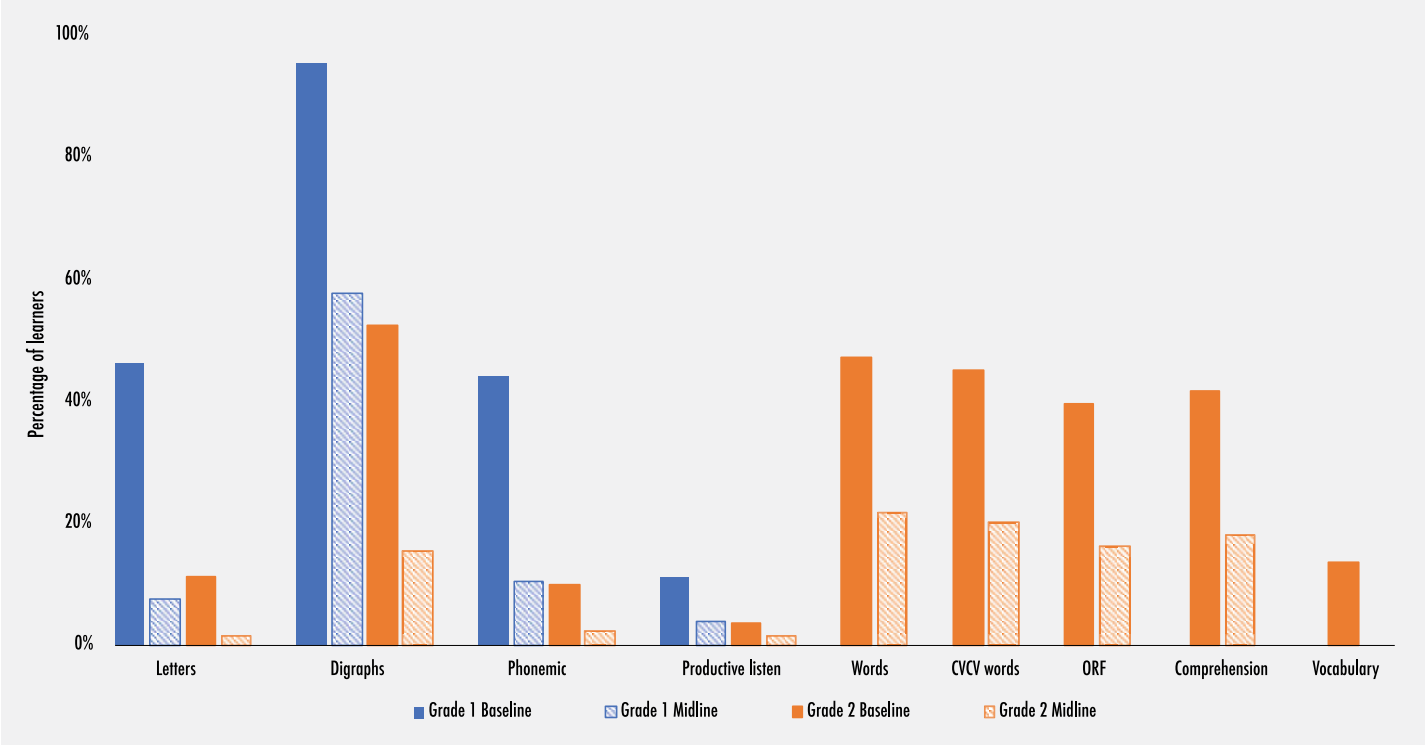
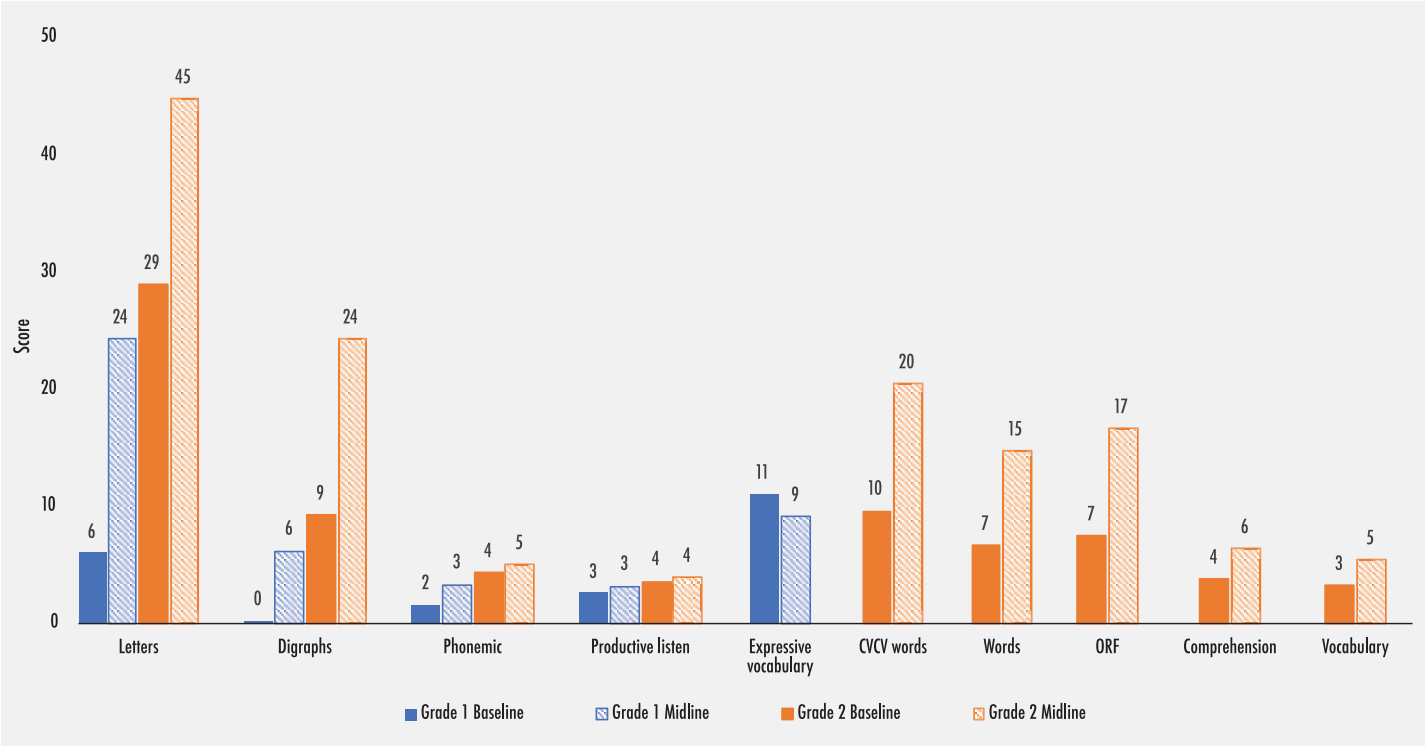


Figure 19. Average scores for control group learners at baseline and midline



7.2. Treatment Effects

7.2.1. PRIMARY OUTCOME MEASURES

In a similar fashion to Cilliers et al. (2019), we constructed a composite score of the isiXhosa reading proficiency based on the different EGRA sub-tasks on which learners were assessed. This was done for both the full sample for the common tasks on which learners in both Grades were assessed, as well as for Grade 2's only (given the additional fluency and comprehension tasks that they were assessed on). The statistical method used, Principal Components Analysis (PCA), reduces the data from the different sub-tasks to create a single variable that captures the most common variation among them: the first principal component. Intuitively, the principal component is taken to be reflective of a common underlying construct, which we here take to reflect isiXhosa reading proficiency. Only the control group's midline sub-tasks are used to construct the index, as these scores give the "business as usual" weighting of the respective factors to the composite reading proficiency index. In order to simplify interpretation, the composite index was standardised by subtracting the control group mean and dividing by its standard deviation (allowing for interpretation in terms of standard deviations).

Intuitively, the principal component is taken to be reflective of a common underlying construct, which we here take to reflect isiXhosa reading proficiency.

The purpose of the composite score is to create one transparent and clearly defined overarching measure of programme impacts. Statistically, it serves as a reassurance that our overall assessment of programme impact, heterogeneous treatment impacts and robustness checks are not selectively reported for certain sub-tasks and/or sub-groups. Nevertheless, given that i) the sub-tests do not all necessarily fit together in one coherent whole and ii) that we are also interested in the impacts of the programme on certain foundational components on the path to reading for meaning, results for the main estimation model are also reported for each sub-task individually.

Given the aim of constructing a reading proficiency index, two tasks were left out of the index: i) the vocabulary task and ii) the productive listening task. The former had had severe ceiling effects (with more than 50 percent of learners scoring full marks) - which affects its usefulness in the index. An exploratory factor analysis indicated that the productive listening task had a low item-rest correlation and loaded higher on the second underlying factor that seems to be indicative of oral literacy skills (and not of reading proficiency) (see Baseline report, 2019: 31). A second composite score was constructed for Grade 2's only. This composite score places a slightly heavier weighting on the reading fluency and -comprehension components skills by construction, as it includes the second reading fluency- and reading comprehension tasks, as well as the sentence choice task that was conducted on Grade 2 learners only. The two reading proficiency scores are very similar, however, with a correlation coefficient between the two scores equal to 0.995.

Given the stated programme objective that all learners should be reading for meaning by the end of Grade 3, reading comprehension measures are also considered as primary outcomes. Following on the discussion of the two main reading

20. Note that a learner was judged able to attempt a comprehension question based on the amount of words that they attempted (i.e. how far into the passage they read, whether or not they read the respective words correctly).lacement teachers, the attrition rate is 16%.

Table 5: How many comprehension questions learners can attempt based on their reading speed

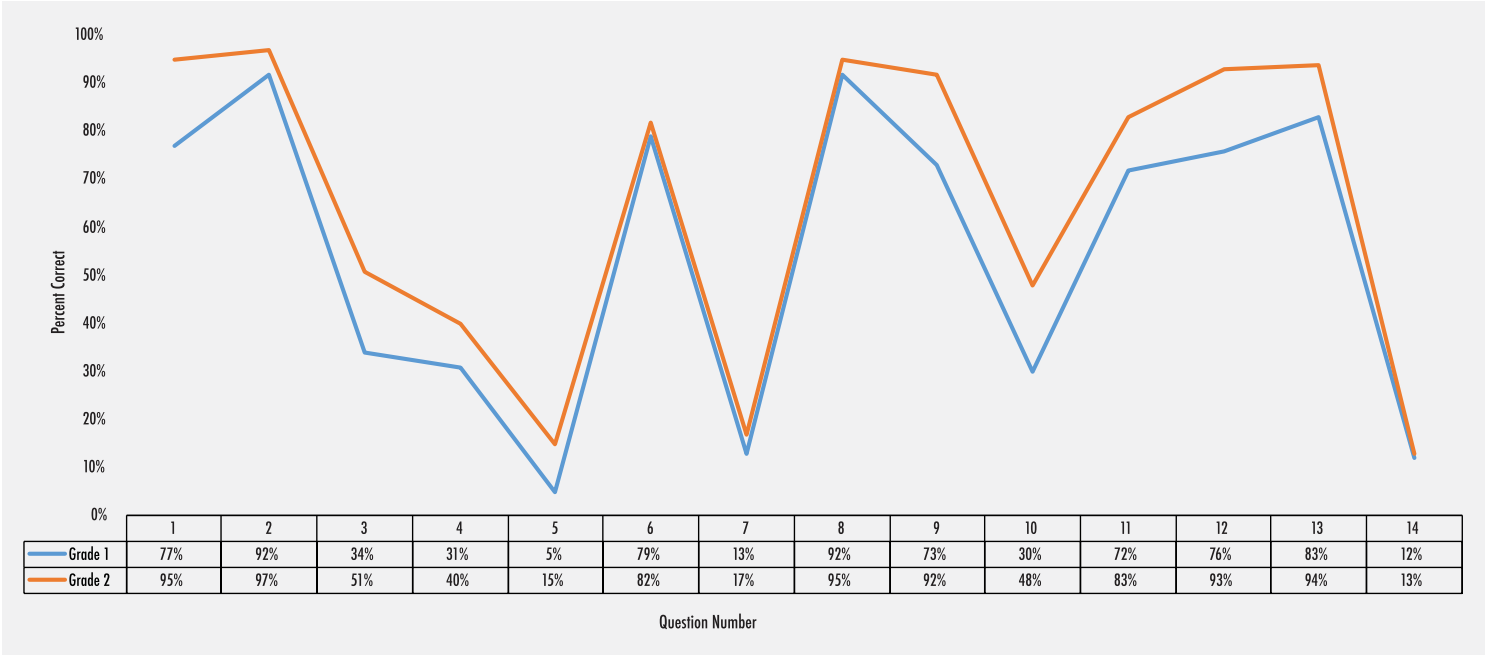
	Can attempt 1st inferential question (>=13 words read)		Can attempt all questions (=41 words read)	
	Grade 1	Grade 2	Grade 1	Grade 2
% after 1 minute	20%	65%	0%	7%
% after 3 minutes	43%	82%	21%	63%

comprehension sub-tasks in section 5.4 above, the additional two minutes of reading time enabled learners to answer a far greater share of the comprehension questions. ***Table 5*** below provides an illustrative example based on the first reading comprehension task, completed by both Grades. For learners to have read far enough to attempt the first inferential comprehension question (the fifth question out of 14), they must have read at least the first 13 words from the passage. Only 20 percent of Grade 1's and 65 percent of Grade 2's managed to read this many words in the first 60 seconds of the reading fluency task. The additional two minutes allowed an additional 23 percent of Grade 1's and 17 percent of Grade 2's to reach the point in the passage where they could attempt the first inferential question. The additional two minutes also allowed a fifth of Grade 1 learners to complete the whole passage, where no Grade 1 could achieve this in only one minute. Almost two thirds of Grade 2's could complete the whole passage after three minutes, up from less than ten percent in one minute.

The average scores on the comprehension tasks are thus slightly misleading to the extent that one is only interested on how well learners were able to answer the comprehension questions which they had read far enough to attempt. For example, the average raw comprehension score for all Grade 2's is 47 percent on the first comprehension task and 41 percent on the second. However, only six in ten Grade 2 learners could attempt all 14 comprehension questions on the first comprehension task, whilst only approximately half (47 percent) of the same group read fast enough to attempt all ten comprehension questions on the second passage. For these two subsets of Grade 2's who attempted all the respective comprehension task questions, their average scores were 65 percent and 67 percent on the two tasks respectively

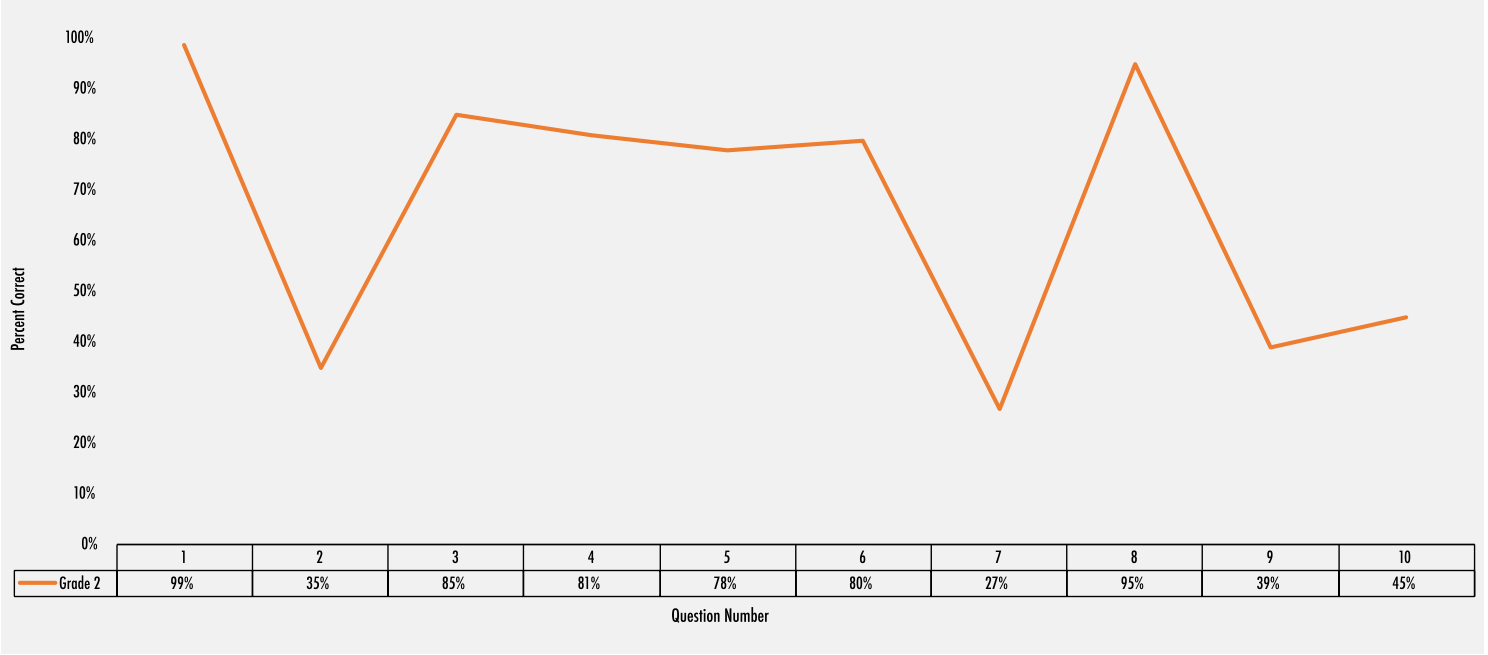
Figures 20 and 21 indicate the percentage of learners who could correctly answer each question of the two paragraph reading comprehension tasks, but only for those learners who finished reading the whole passage in three minutes. There is a large variability in learners' ability to answer the questions across both reading comprehension tasks. In particular, learners fared the worst in the interpretive and inferential questions in the first comprehensions task (questions 5, 7, and 14). For the second comprehension task, learners also scored very low on three of the interpretive questions (questions 2, 9 and 10), whilst the other lowest scoring items (question 7) was a factual detail from the story that learners generally struggled to recall.

Figure 20: How learners who attempted all questions fared by question (Reading Comprehension 1, by Grade)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Grade 1	77%	92%	34%	31%	5%	79%	13%	92%	73%	30%	72%	76%	83%	12%
Grade 2	95%	97%	51%	40%	15%	82%	17%	95%	92%	48%	83%	93%	94%	13%

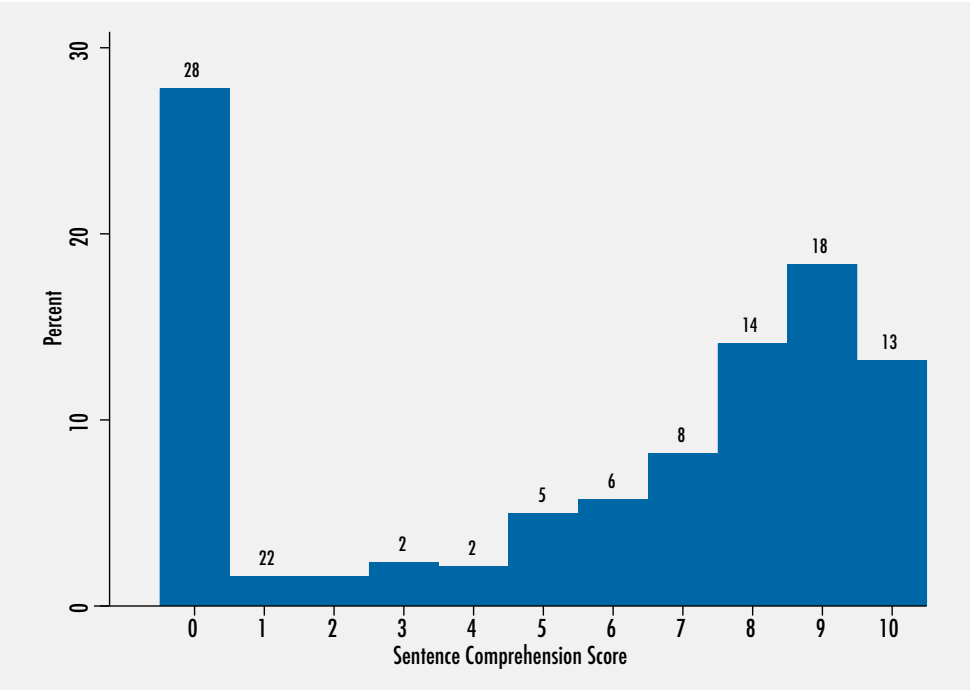
Figure 21: How learners who attempted all questions fared by question (Reading Comprehension II, Grade 2 only)



	1	2	3	4	5	6	7	8	9	10
Grade 2	99%	35%	85%	81%	78%	80%	27%	95%	39%	45%

Figure 22 shows the distribution of how Grade 2 leaners scored on the sentence comprehension task. The task was included to obtain a measurement of reading comprehension that was less contingent on learners' reading fluency. However, the scores for the task are clustered at both the bottom (28 percent score zero) and top (46 percent score 80 percent or more) of the distribution. One in five learners could not read the first three sentences and the task was discontinued. These learners make up the vast majority of the learners scoring zero on the task. Almost all (91percent) of the learners who did not attempt a reading comprehension question, scored zero on the sentence comprehension task. This suggests that the task is somewhat limited in the extent to which it can discriminate between learners reading comprehension ability at both the lower and upper ends of the distribution.

Figure 22: Histogram of Sentence Comprehension Scores (Grade 2 only)



7.2.2. DESCRIPITIVE ANALYSIS OF INTERVENTION IMPACTS

The distribution of the midline composite score is shown separately by treatment status in ***Figure 23.*** For each level of the composite score, the lines indicate the proportion of learners with scores at that level or greater. Intervention school learners appear to outperform those in the control schools across the distribution. For example, around 47 percent of treatment learners have a composite score above zero, in contrast to 42 percent of control learners. Comparing learners at the same point in the two groups' respective distributions of midline reading proficiency scores, learners in the intervention group score 0.06 s.d. higher at the 25th percentile, 0.23 s.d. higher at the median, and 0.25 s.d. higher at the 75th percentile.

We next turn to the shifts in the distributions of selected Grade relevant task. The histograms in ***figures 24 to 26*** indicate i) how learners' reading abilities in the control group improved in the status quo learning environment over one year of schooling and ii) to what extent the intervention further shifted the distribution of learner outcomes in addition to what occurred in the comparison group .

In each case the left panel displays the shift in the control group from the baseline- to midline assessment, whilst the right panel overlays the distribution of learner outcomes of the control- and intervention school learners at midline. The orange filled bars represent control school learners' outcomes at baseline; the orange outlined, unfilled bars represent control learners at midline (i.e. the same in both left and right panels); and the blue filled bars represent intervention school learners at midline.

Figure 24 shows the overlaid histograms of the percentage of Grade 1's who could correctly identify the number of letter sounds per minute corresponding to the respective five-letter bins. For example, at baseline 64 percent of Grade 1 control school learners could only read between zero and five correct letter sounds per minute. At midline, the share of control school learners whose scores fell in this interval decreased to 16 percent. For intervention school learners, however, the share of learners who could identify five or less correct letter sounds per minute was even less - only 12 percent. Overall, ***Figure 24*** provides a sense graphically of the extent to which the intervention further shifted

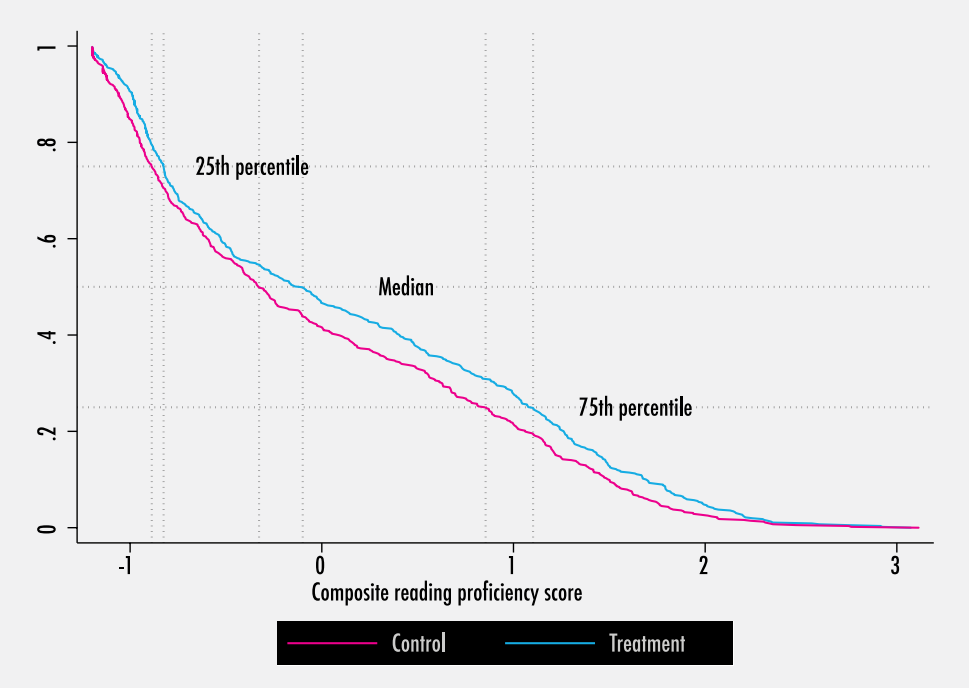
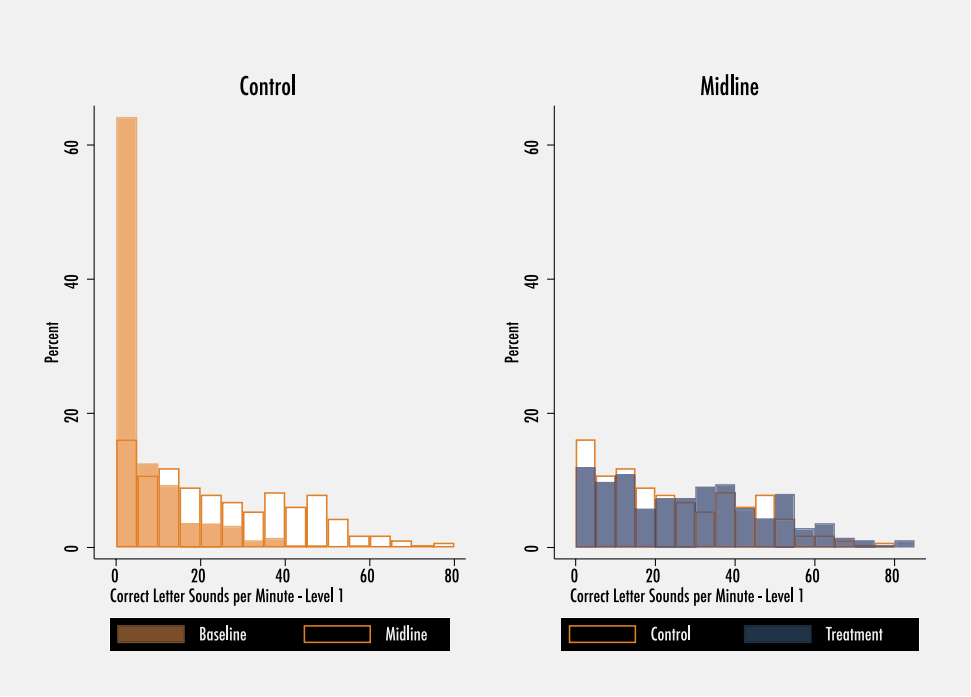


Figure 23: Distribution of midline composite scores by treatment status (full sample of both Grades combined)

21. Summary information on the distribution of each sub-task by Grade is displayed in Appendix table A4.

Figure 24: Histograms of shifts in Grade 1 letter sound recognition ability (bin-width= 5 letter sounds)

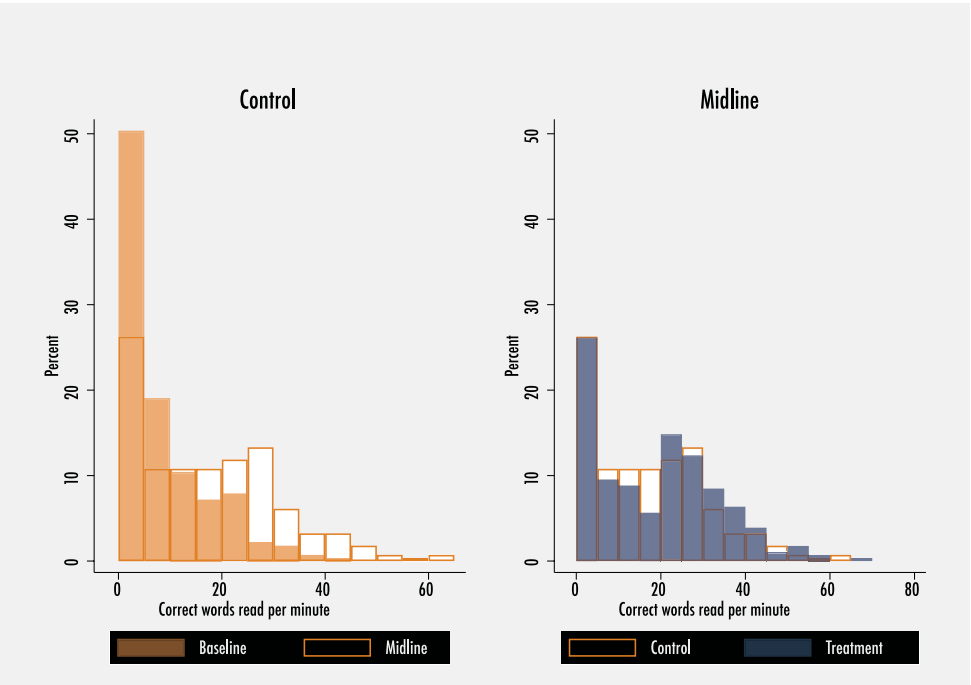


the distribution of Grade 1's letter recognition scores to the right, and how significant this is relative to the growth that that occurred under status quo learning environments.

Figure 25 and 26 show similar rightward shifts for Grade 2 learners on higher order reading proficiency tasks: i) oral reading fluency and ii) reading comprehension. From the left panel in ***Figure 25***, it is evident that there was a significant decrease over the year in the share of control school Grade 2 learners with an oral reading fluency of ten or less correct words per minute. For these same control school Grade 2's there is an increase in share of learners who could read 16 correct words per minute or more at midline. In turn, intervention school learners saw their distribution shift even further to the right, with a relatively greater share of learners scoring above 20 correct words per minute at midline.

For Grade 2 learners' reading

Figure 25: Histograms of shifts in Grade 2 oral reading fluency ability (bin-width= 5 words)



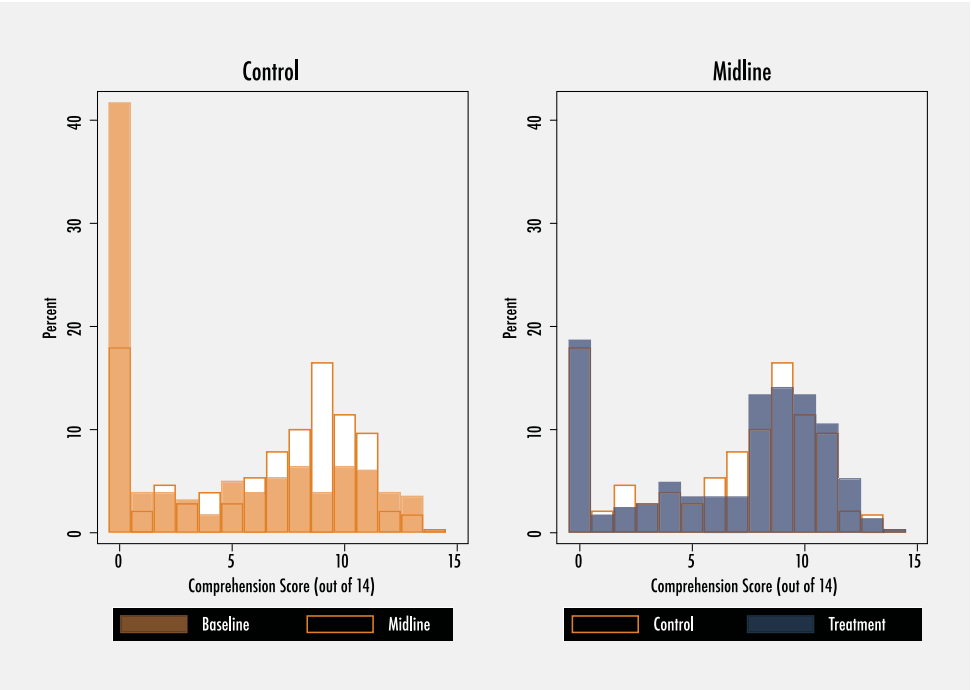
comprehension score, ***Figure 26*** indicates that there were significantly less learners in control schools scoring zero on the comprehension task at midline, in conjunction with a much greater proportion of these learners scoring six or more (out of 14) on the comprehension task. In comparison to the control school Grade 2's, intervention school learners had a higher share of learners with scores of eight or more on the comprehension assessment at midline. In sum, ***Figures 25 and 26*** graphically illustrate two themes with respect to the shifts in the distributions of Grade 2 learners' scores in these higher order domains of reading proficiency. First, the status quo schooling progression over the year results in a significant decrease in the proportion of Grade 2 learners scoring in and around zero on these tasks. Second, intervention schools have a greater share of learners obtaining relatively high scores on these tasks at midline.

A simple comparison of means on the sub-tasks provides an initial indication of programme impacts (see Appendix ***Table A4*** for detailed data on the means and distributions of each sub task, by Grade). ***Figure 27*** shows the difference in mean scores between treatment and control learners. For both the common assessments conducted on learners in both Grades, as well as Grade specific assessments, learners in the treatment group almost universally scored higher on average. For example, learners in the treatment group could read four more letters- and three more digraphs and trigraphs per minute than learners in the control group. Similarly, intervention school learners could read between one and two words more per minute than their peers in control schools in the word recognition tasks and the paragraph reading fluency task, respectively. The only exception is for the vocabulary sub-task where scores in both groups suffered from high ceiling effects. In the next section, we investigate whether these differences are statistically significant.

7.2.3. REGRESSION ANALYSIS OF INTERVENTION IMPACTS

Random assignment of schools to intervention and control groups, and the fact that the two groups were still balanced at midline, ensures that a simple comparison of means across

Figure 26: Histograms of shifts in Grade 2 reading comprehension ability (bin-width= 1 comprehension score)



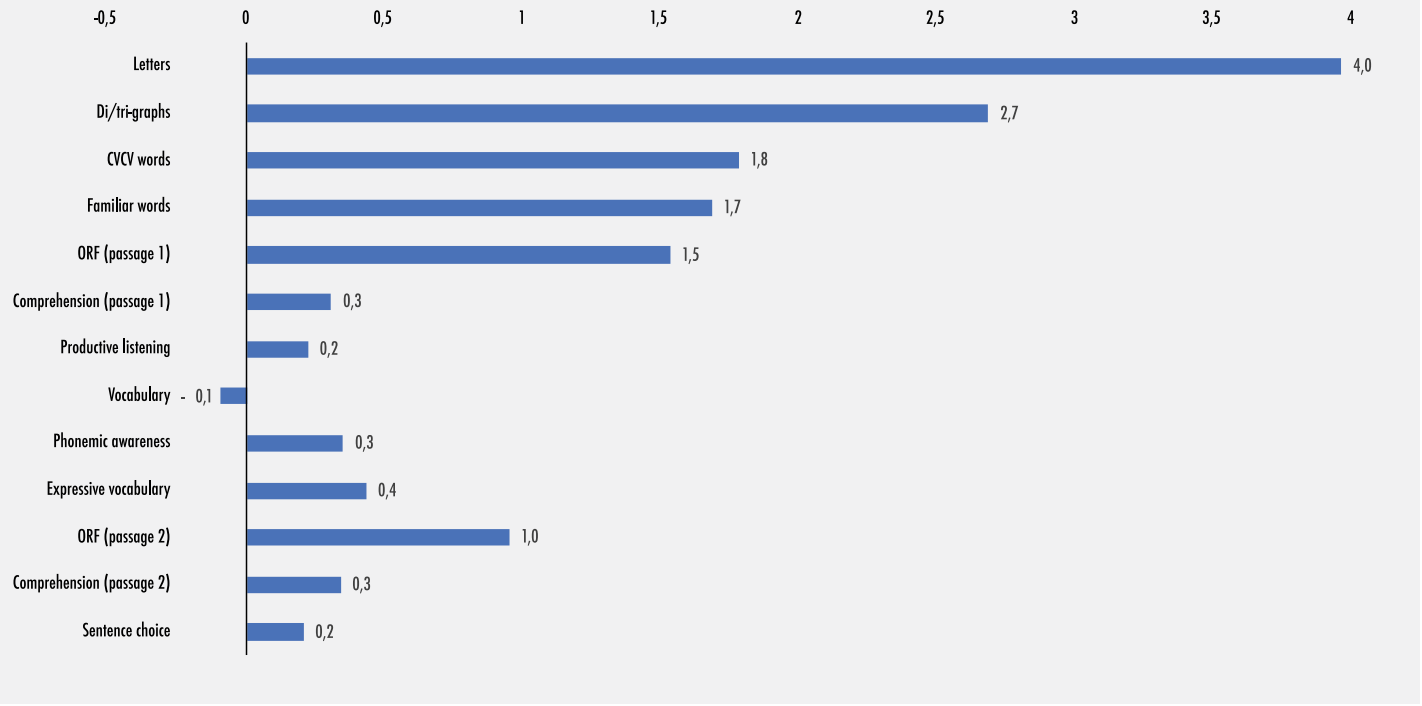


Figure 27. Differences between treatment and control means on all sub-tasks (treatment minus control)

learners in the intervention and control schools provides a reliable estimate of the programme impacts. However, regression analysis of the programme impacts allows one to i) control for any incidental pre-randomization differences between the two groups, ii) account for non-random attrition and iii) increases the precision of the estimates by including variables that explain a large share of the variation in outcomes that are unrelated to the intervention. All results reported therefore control separately for each relevant measure of reading proficiency collected at baseline[22], learner level characteristics, and household assets, as well as strata fixed effects. In the cases where learners had missing data on a certain dimension of the control variables (say, if they did not answer a question on some household asset), a missing value was imputed and a separate dummy variable was included to indicate missingness as a control[23]. Since schools (i.e. clusters of learners) were randomly assigned to either the intervention or control groups, and not the individual learners, it is best practice to cluster standard errors at the level of randomisation (Abadie, Athey, Imbens and Wooldrige, 2017)[24].

The impact of the programme on reading proficiency for the full cohort of learners in intervention schools is 0.17 standard deviations (s.d.) over one year of exposure to the programme (see the composite score estimate in ***Figure 28*** below). Appendix ***Table A3*** reports the effect size point estimates in standard deviations, standard errors of the estimates, the regression estimated p-values, as well as the randomisation-based inference constructed p-values (as recommended by Athey and Imbens, 2017) for all the results that follow.

To get a sense of the relative size of the impacts on each sub-task and how these relate to the impact on the composite reading proficiency measure, ***Figure 28*** below reports the estimated effect sizes on standardised versions of the various sub-tasks and composite scores. The exercise is repeated for each Grade, including the Grade-specific controls. The darker shaded areas of the bars display the 90 percent confidence interval, and the lighter shaded fringes end at the 95 percent confidence intervals for the estimated

22. In cases where the analysis is done on the full sample, this includes the common tasks assessed for both Grade 1's and Grade 2's at baseline. For analyses on one Grade only, the relevant Grade specific assessments at baseline were included as additional controls. See Table 2 for the layout of common- and Grade specific tasks at baseline.

23. Missing values were assigned a value of zero if the variable is categorical, whilst missing observations on continuous variables were set equal to the sample mean (in a similar fashion to Cilliers et al., 2019).

24. Intuitively, two learners drawn at random from within the same school are more likely to have similar reading outcomes than two learners from the full sample. One expects that the school where learners find themselves explains some of the variation in their reading outcomes because i) learners within the same school are relatively more similar in terms of their daily circumstances and a range of unobservable characteristics that are important for reading outcomes and ii) learners in the same school have the same teachers, access to the same reading materials, are affected by the same school levels shocks (like a principal resigning), etc. – all of which impact on their reading outcomes. When indicating the degree of certainty in estimates of the programme impact, cluster/group level standard errors take into account that for learners within each school there is some degree variation in their reading outcomes that is not explained by the model, but which is specific to the school in which the learners are. For an extended, accessible discussion on the group-level standard errors in cluster RCTs, see Glennerster and Takavarasha (2013: 356-361).

impacts.

Figure 28 shows the estimated impacts for the full sample. The estimated impacts were relatively large and consistent across the range of emergent- and early literacy tasks. The largest point estimate of programme impact on any one sub-task was on the correct letter sounds per minute task (0.22 s.d.). The point estimates of impacts on learners' ability to correctly identify more complex letter sounds (digraphs and trigraphs), to identify and manipulate phonemes (phonemic awareness), or to correctly answer questions based on a passage read aloud to them (productive listening comprehension) were all 0.16 s.d. or larger. The impacts on word recognition and paragraph reading fluency tasks ranged between 0.14 s.d. and 0.17 s.d.'s. Lastly, the effect estimated on the vocabulary task is very noisy. This is to be expected, given that learners generally scored at or close to the maximum on this task and it was therefore unable to discriminate between learners. The vocabulary task is therefore excluded from subsequent analyses, and was included here only for transparency purposes. For the end-line assessment, a more challenging vocabulary assessment will need to be included.

Reading comprehension, as the main outcome of interest of the programme, had a positive point estimate (0.11 s.d.) that was not statistically significant (see ***Figure 28*** above, as well as the p-values reported in appendix ***Table A3***). A smaller effect on reading comprehension at this stage accords with the theory of a sequential acquisition of literacy skills. Foundational components, like letter-sound knowledge, phonemic awareness, and word recognition are prerequisites for progressing to reading passages. Learners arguably require a range of foundational literacy abilities before they can read with some level of speed and accuracy (i.e. fluency), and in turn, then need to read with a certain minimum level of fluency in order to comprehend what they are reading[25].Nevertheless, the impact of the programme on reading comprehension and the extent to which literacy skills are hierarchical will become clearer after the next round of data collection. For now, we look at results by Grade to gain some insight on the literacy development trajectories of learners in intervention and control schools.

Figure 29 displays the estimates

25. See discussion in Spaull, Pretorius and Mohohlwane (2020: 5-8) for a discussion of the hierarchical nature of language acquisition and its applicability to learning to read African languages in the South African context.

26. The less precise estimate for the Grade 1 specific treatment impacts follows not only from the smaller sample size of the group (relative to the full sample), but also from the fact that the limited literacy and pre-literacy sub-tasks on which Grade 1's were assessed at the beginning of the year do not account for as large a share of the variation in their reading outcomes at midline. There are also a higher percentage of Grade 1 learners scoring zero on the various sub-tasks resulting in lower variation between learners on the reading proficiency measure than for Grade 2 learners.

27. The original passage consisted of 41 words and had a descriptive picture. The newly introduced passage consisted of 55 words and had no descriptive picture accompanying it.

Figure 28: Treatment effects for common tasks assessed on both Grades, overall and by sub- task

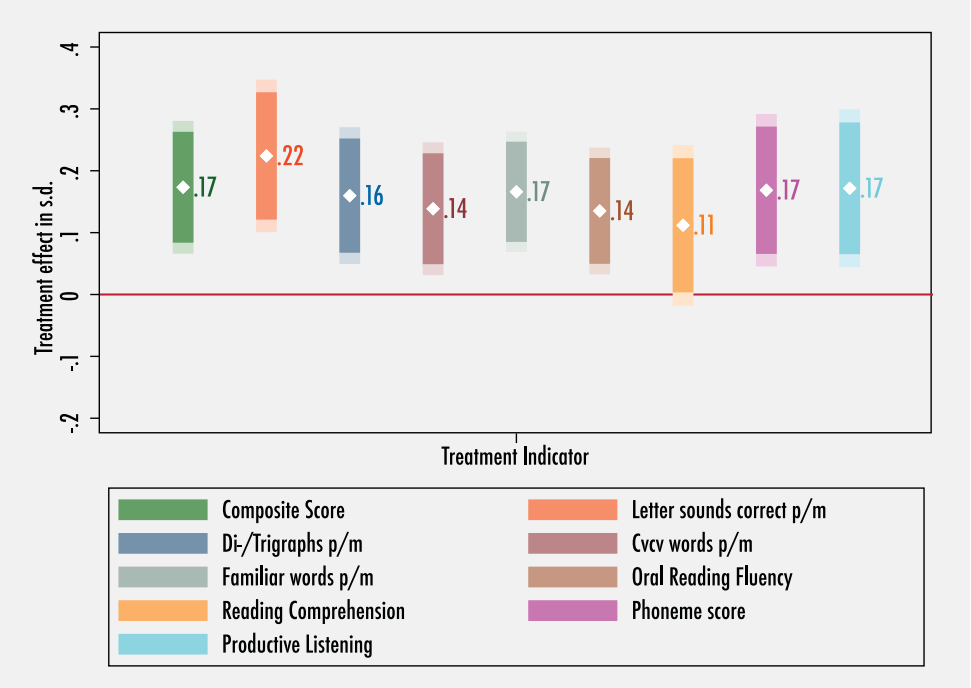
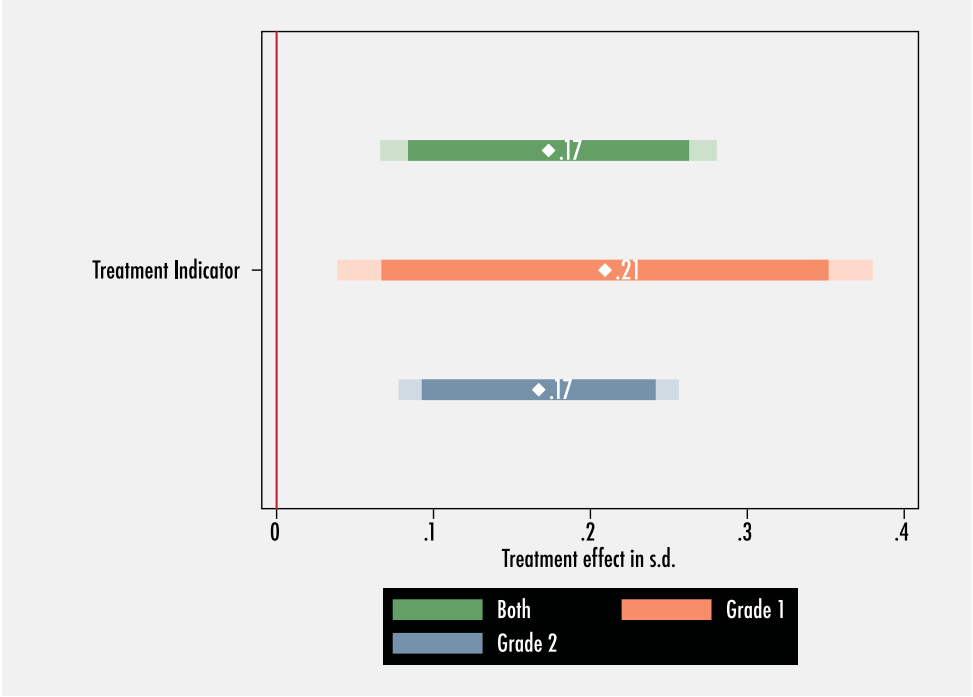
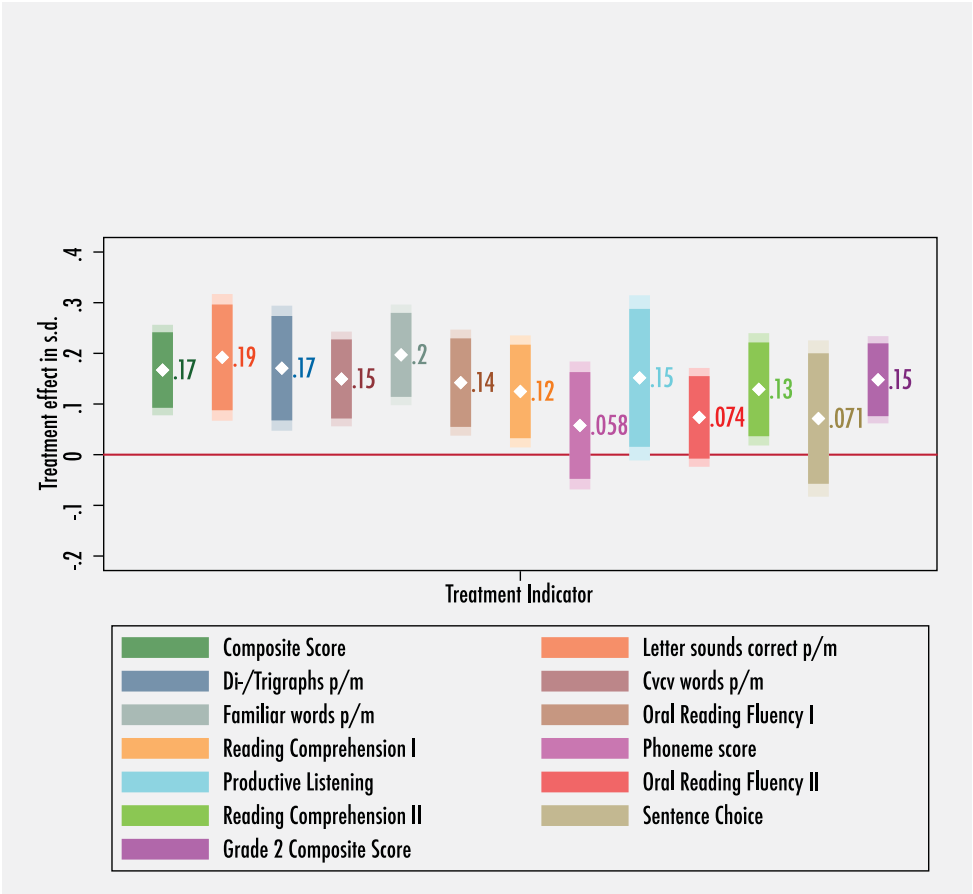


Figure 29: Treatment effects on reading proficiency, for full sample and by Grade



of programme impact on Grade 1 and Grade 2 reading proficiency separately. The effect on Grade 1 reading proficiency is estimated to be a relatively larger 0.21 s.d., but also with a larger standard error[26]. Controlling for additional Grade 2 specific baseline tasks and re-estimating the model for Grade 2 learners yields an effect size that is slightly lower at 0.16 s.d., but more precisely estimated. However, given the overlapping confidence intervals, we cannot infer that the programme

Figure 30: Treatment effects for Grade 2's only, overall and by sub-task



impact was larger on any one Grade specifically.

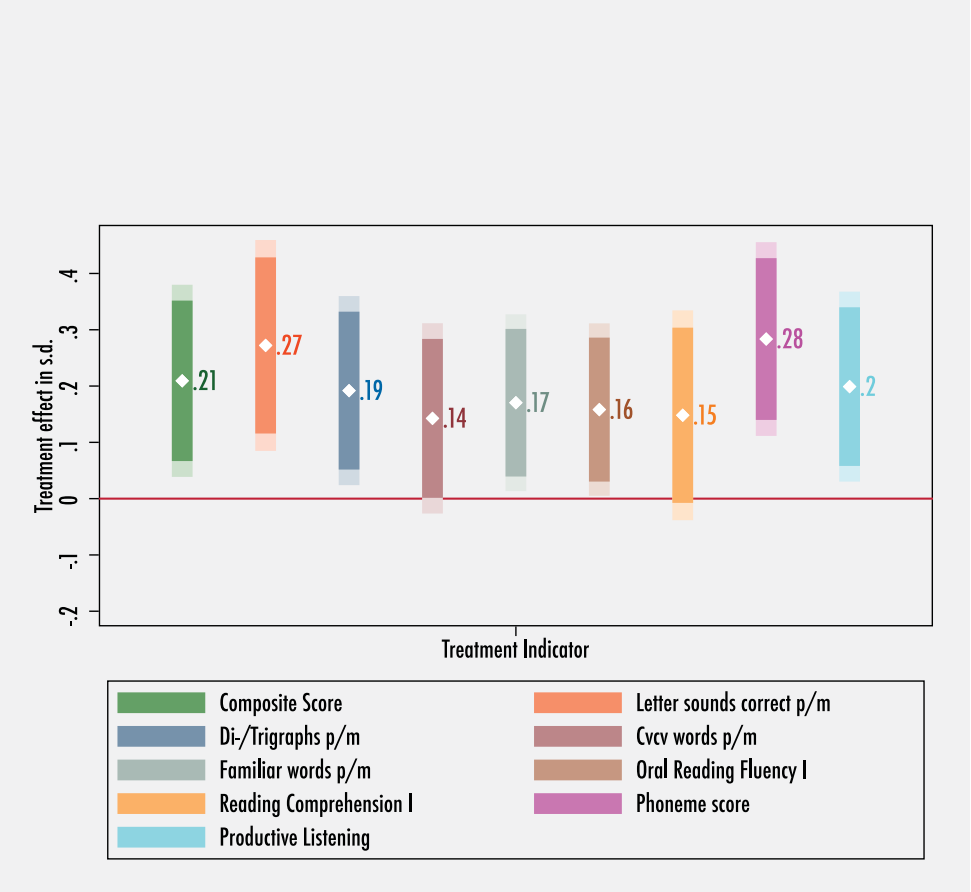
Using the Grade 2 specific reading proficiency outcome measure, which weighs more heavily reading fluency and comprehension skills, yields only a slightly lower estimate of the intervention's impact: 0.15 s.d (see Grade 2 specific ***Figure 30***). ***Figure 30*** also reports programme impacts on each sub-task, estimated on the Grade 2 sample only. Programme impacts again seem to be largest on certain foundational literacy skills for Grade 2 learners: identifying letter sounds (0.19 and 0.17 s.d.'s respectively) and word recognition (0.15 and 0.19 s.d.'s). Despite the smaller sample size, estimates of the intervention impacts are estimated with a relatively high amount of precision - largely due to the extent to which baseline reading proficiency measures for Grade 2 learners explain their midline literacy outcomes. Estimated effects on phonemic awareness and productive listening are noisier, however. The point estimate of the effect size on the productive listening is comparatively large (0.15 s.d.), but the effects on these two tasks are not statistically distinguishable from zero.

Grade 2 learners' reading fluency was assessed on two separate passages at midline. The first passage was the same passage used to assess reading fluency and -comprehension for Grade 2 learners at baseline, whilst the newly introduced second passage was slightly longer and more challenging[27]. The effect of the programme on the reading fluency for Grade 2's on the original, simpler passage was 0.14 s.d, in contrast to the smaller and not significant effect of 0.05 s.d. on the new, more challenging passage.

Effects on the reading comprehension are consistent with the notion that a certain level of emergent and pre-literacy skills must be acquired before seeing shifts on these higher order outcomes. In contrast to the full sample, there are statistically significant effects

28. Given that the task provided learners with a binary yes/no choice as to whether the short preceding sentence pair which they read made sense, it might be that learners who did not understand the passage would have resorted to guessing the answers. If all learners guessed all items, that average score would be 2.5. Looking at the distribution of scores it is clear that the majority of learners are clearly not just guessing (refer to figure 3 earlier). Nevertheless, any guessing would introduce some noise into the measure. As discussed in section 7.2.1., however, the task seems be limited in its ability to differentiate among learners' underlying reading comprehension ability - given the extent to which learners' scores on the task cluster at either the bottom or top of the distribution

Figure 31: Treatment effects for Grade 1 sample, overall and by sub-task



Programme impacts for Grade 1 learners are largest on emergent- and pre-literacy skills: letter sound recognition.

of the programme on reading comprehension for both passages. Grade 2 learners in treatment schools score around 0.12 to 0.13 s.d. higher than their peers in control schools. The estimate on the impact on the additional sentence comprehension task was smaller and less precisely estimated. This smaller estimated effect size and the noise in the estimate might well result from the nature of the task itself[28]. Considered overall, the programme significantly shifts reading comprehension outcomes at the Grade 2 level.

Programme impacts for Grade 1 learners are largest on emergent- and pre-literacy skills: letter sound recognition (0.27 s.d. for single letters and a slightly lower 0.19 s.d. for more complex digraphs and trigraphs), phonemic awareness (0.28 s.d.) and productive listening comprehension (0.20 s.d.) (***Figure 31***). It is interesting to note the relatively large difference in effect sizes on phonemic awareness for Grade 1 and Grade 2 learners, with a 0.23 s.d. difference in point estimates. Viewed alongside the relatively larger point estimate for Grade 1's on another first order literacy component, simple letter sound recognition, these results would be consistent with the idea that letter sound knowledge and the ability to manipulate phonemes are important foundational skills required for learners to sound out novel words and better progress toward word reading (Spaull et al., 2020: 5-6). Declining impacts on these two tasks over time would be in line with broader findings that alphabetic awareness has a narrow developmental window (Ouelette and Haly, 2013). If Grade 1 learners in the intervention schools see the greatest impact on their ability to link letters to sounds and to become more aware of how these letter sounds function within words, we would expect them to be better able to decode new words at the subsequent word- and sentence reading stages of their developmental

process. In turn, we would expect diminishing returns and smaller effects (relative to generally lagging control school learners) in subsequent assessments on these previously acquired alphabetic awareness skills. It will be interesting to observe if these patterns do indeed play out over time for the 2019 Grade 1 cohort.

Effect sizes on the two respective word recognition tasks are 0.14 s.d. and 0.17 s.d., but the point estimate on the former CVCV word reading task is only significant at the p=0.1 level[29]. The programme impact on Grade 1 learners reading fluency is both positive and significant (0.16 s.d.). Finally, the point estimate reading comprehension effect is positive, but it is not significant (both the 90% and 95% confidence intervals overlap with the zero-line). Overall, the point estimates on the word recognition, reading fluency and comprehension tasks are in line with those estimated for Grade 2 learners. However, the effects are more noisily estimated because i) Grade 1 learners could not be assessed on these higher order literacy skills right at the start of their schooling career ,ii) there were floor effects on simpler baseline tasks that would have predicted midline word- and paragraph reading outcomes (like the letter sound recognition tasks), and iii) the fact that more than half the Grade 1 learners assessed at midline scored zero on each of the aforementioned tasks[30].

Whilst impact estimates measured in standard deviations are useful in providing a relative sense of the size on programme effects on various sub-tasks, they are not very intuitive and give us less of a sense of what learning gains translate to in practice. One way to better gauge the practical significance of learning effects is to interpret them relative to a year of learning in the control group. This provides us with an estimate of how large the additional learning gains in the intervention group are relative to the "business as usual" learning gains that accrued over the academic year to control group learners. This measurement requires a sub-task to have been assessed on the relevant Grade at both the baseline and midline assessments, however, and is therefore only reported for the tasks where this was the case[31].

For each sub-task conducted on Grade 2 learners, ***Table 6*** below displays for the control group the mean outcome at baseline, the growth (or difference in means) between baseline and midline, the standard deviations of the respective estimates (to get a sense of their precision), and the number of observations on which the estimates are based. The second part of the table provides the estimate of the effect size on learners in intervention schools (also in task-specific units, like words read correctly per minute) and reinterprets this as a percentage of the learning that took place under status quo conditions in control schools over the academic year. Results are reported only for those subtasks on which the programme effect was significant at the p=0.05 level.

At baseline, Grade 2 learners in control schools could identify 29

29. This is with reference to the regression-based estimate of the p-value. The randomization inference-based p-value (p=0.16) implies that the effect on Grade 1 learners' ability to correctly recognize CVCV words is not statistically significant.

30. For the CVCV and familiar word tasks respectively, 55% and 52% of Grade 1's scored zero at midline. In turn, the share of learners scoring zero on reading fluency and -comprehension was 56% and 57%.

31. As noted earlier, the Vocabulary task is not discussed due to extreme ceiling effects.

32. Note that all interpretations in this section are for mean outcomes, unless clearly stated otherwise.

33. Note that of these learners, almost 60 percent could still not correctly identify a single digraph or trigraph by the end of one year of schooling.

Table 6: Treatment effects in terms of a year of learning, Grade 2

	Control Group				Treat. v Control	
	Baseline Mean (units)	Baseline Standard Deviation	Growth in mean	N	Effect size (units)	% of year of learning
Letter sounds	29,0	19,4	15,8	278	4,2	27%
Digraphs and Trigraphs	9,3	13,8	15,0	278	3,0	20%
CVCV Words	9,6	11,9	10,8	278	2,2	20%
Familiar Words	6,7	8,6	8,0	278	2,1	26%
Oral Reading Fluency	7,5	9,1	9,2	278	1,8	19%
Reading Comprehension	4,3	4,6	2,1	278	0,5	24%

Grade 2 learners in intervention schools could therefore expect to see their word- andparagraph reading skills improve by between a fifth and a quarter more than what took place in schools wherethe intervention did not take place.

correct letter sounds per minute[32]. This grew to 45 letter sounds per minute at midline (see the third column in ***Table 6***). The estimated impact of four additional correct letter sounds per minute (fifth column, ***Table 6***) for Grade 2 learners in the intervention schools thus equates to 27 percent of a year, or roughly a school term's worth, of learning. For the same group of learners' ability to correctly identify digraphs and trigraphs, a treatment effect of 3 additional letter sounds per minute translates to one fifth of what learners in control schools gained over the year Therefore, even though effect sizes like "three to four more correct letter sounds per minute" might not give the impression of large learning gains when one does not have a sense of the context, these gains are indeed considerable when viewed relative to the status quo learning gains accrued over a full academic year.

For the word reading and reading fluency tasks, effect sizes of approximately two additional words correctly identified per minute translate to 20 and 26 percent of a year of learning in control schools, for CVCV and familiar word reading respectively. In control schools, Grade 2 learners read less than eight words correctly per minute from a short passage at baseline. This grew to almost 17 correct words per minute at midline, which implies that the two additional words per minute treatment effect represents a fifth of a year's worth of learning for reading fluency. Grade 2 learners in intervention schools could therefore expect to see their word- and paragraph reading skills improve by between a fifth and a quarter more than what took place in schools where the intervention did not take place.

At midline, Grade 2 control school learners answered 46 percent of the comprehension task questions correctly, up from only 31 percent at baseline. The intervention impact of 4 percentage points higher comprehension scores (0.5 more questions answered correctly out of a total of 14) by intervention school Grade 2 learners translates to 24 percent of a year, or roughly a term, of learning. This provides one with a greater appreciation of the size of the impact on reading comprehension, to complement the 0.14 s.d. effect size

Table 7: Treatment effects in terms of a year of learning, Grade 1

	Control Group				Treatment v. Control	
	Baseline Mean (units)	Baseline Standard Deviation	Growth in Mean	N	Effect Size (units)	% of year of learning
Letter Sounds	6,0	9,6	18,3	279	6,0	33%
Digraphs and Trigraphs	0,2	1,4	6,0	279	3,5	58%
Phonemic Awareness	1,6	1,8	1,7	279	0,6	36%
Productive Listening	2,6	1,5	0,6	279	0,3	46%

on reading comprehension for Grade 2's displayed in ***Figure 30*** earlier. Overall, reading comprehension gains for Grade 2's in intervention schools seem promising and are potentially large relative to status quo levels of learning.

For Grade 1 learners, only four of the tasks that were assessed at both baseline and midline had significant effect sizes (see ***Table 7*** below). These tasks were all emergent- and pre-literacy tasks, and on each one the effect size was between a third and two-thirds of a year's worth of learning in control schools.

By the end of Grade 1, control school learners more than quadrupled the amount of simple letter sounds that they could identify correctly per minute (from six to 24). The estimated six additional correct letter sounds identified by intervention school learners is thus large both relative to what was gained under the business as usual Grade 1 schooling environment, and with respect to the four letter sounds per minute treatment effect for Grade 2 learners. For the more complex digraphs and trigraphs, 96 percent of all Grade 1 learners could not identify a single letter sound at baseline. The average digraphs and trigraphs correctly identified per minute moved from a base of almost zero to six at midline for control school learners[33]. The effect size of almost four correct letter sounds per minute thus translates to more than a half of a year's worth of progress under control school conditions. This is closer to the effect of the intervention on Grade 2 learners in absolute terms (three to four letter sounds), but almost three times larger when viewed in terms of the amount of learning that occurred in the Grade specific comparison group.

For the phonemic awareness task, control school Grade 1's basically doubled their mark scored to three (out of ten) over the academic year. In comparison, the impact estimate for intervention school learners is equal to another third of a year of learning over and above what occurred in the comparison group. The treatment effect was also large on the productive listening comprehension task (counting out of six), where a 0.4 mark treatment effect translates to 46 percent of a year of learning. Considered collectively, the larger effects on the four foundational literacy skills in ***Table 7*** for Grade 1 intervention school learners (relative to intervention group Grade 2's) provide further suggestive evidence that literacy skill acquisition might be sequential in nature. Effects on foundational skills seem to diminish as learners reach a certain level of competency, which in turn allows them to move on to higher order decoding, reading and comprehension skills development.

The analysis up to this point has been focussed on the growth in mean outcomes of learners in the two groups. The analysis now turns to how much individual learners improved (or gained) between baseline and midline in their performance for selected Grade relevant subtasks. Focusing on the distribution of gains, learners within treatment and control schools are classified into quartiles. ***Tables 8 to 10*** compare the gains across treatment and control learners by quartiles of the distribution of gains for selected tasks. This also allows one to compare whether the distribution lof earners' reading proficiency gains differs by intervention status. In each case, the first column displays the average gains in the task for control group learners by quartiles of how much

Table 8: Gains in correct letter sounds per minute by treatment status (Grade 1 only)

Quartiles of gains	Control gain	Treatment gain	T-C
Bottom 25%	1,0 (4,2)	4,7 (4,5)	3,7
26th -50th percentiles	10,6 (3,1)	15,6 (3,1)	5,0
51st-75th percentiles	22,8 (3,9)	27,5 (3,7)	4,7
Top 25%	40,7 (8,6)	47,5 (10,2)	6,8
Total	18,3 (15,8)	23,4 (17,2)	5,1

On average, the relative improvement of the intervention school learners over gains in the comparison group is an additional 5 letter sounds per minute.

learners gained/improved between baseline and midline. The second column shows average gains by quartile for treatment group learners. Standard deviations are reported in brackets below the respective means. The third column shows the difference in learning gains between learners in control and intervention by quartiles of learning improvements.

Table 8 indicates the extent to which Grade 1's in different quartiles of the distribution of letter sounds gained compare between the control and intervention schools. For example, the bottom 25 percent of "gainers" in control schools only improved their letter sounds scores by 1 letter sound on average over the year. In the intervention schools, the slowest improving 25 percent of Grade 1's improved at a rate of approximately four more letter sounds than the comparison group over the year. Comparing the two fastest growing quartiles of the two groups, the fastest improving 25 percent of learners in control schools improved by 41 letter sounds over the year, seven letter sounds less than the 48 letter sounds gained by the fastest improving intervention school learners. On average, the relative improvement of the intervention school learners over gains in the comparison group is an additional 5 letter sounds per minute. This difference between the groups is fairly constant across the distribution of learning gains, but it is slightly lower for those learners in the bottom 25 percent of letter sound reading improvement (4 letter sounds). The difference between groups is the largest when comparing the rate of gains at the top 25 percent, where there is a seven letter sound difference in the rate of improvement.

A similar pattern holds for the distribution of improvements in reading comprehension scores between Grade 2 learners in the two groups. In both groups the bottom quartiles of learners in terms reading comprehension score improvements, saw their scores decrease by one correct answer over the year. For the two groups' respective top quartiles in terms of reading comprehension gains, intervention school gainers improved their comprehension scores by 8 correct answers over the year – one more than what the top gainers in control group managed.

Overall, based on the improvement in reading fluency and comprehension outcomes over the year, Grade 2 learners in intervention schools saw their reading proficiency outcomes improve at a faster rate on average than learners in control schools. The intervention school gainers' additional improvement over outcomes in control schools were largest amongst those learners in the

Table 9: Gains in oral reading fluency by treatment status (Grade 2 only)

Quartiles of gains	Control gain	Treatment gain	T-C
Bottom 25%	-0,5 (2,0)	-0,6 (3,0)	-0,1
26th -50th percentiles	4,7 (1,5)	5,3 (2,4)	0,6
51st-75th percentiles	12,0 (2,7)	13,6 (2,3)	1,6
Top 25%	22,1 (4,6)	24,5 (4,9)	2,4
Total	9,2 (8,8)	10,5 (17,2)	1,3

Overall, based on the improvement in reading fluency and comprehension outcomes over the year, Grade 2 learners in intervention schools saw their reading proficiency outcomes improve at a faster rate on average than learners in control schools.

intervention schools who improved the most. The largest differences between the two groups' respective gainers were in the upper half of the distributions of improvements.

Table 9 similarly displays the gains for Grade 2 control and intervention group learners, by quartile of improvement in oral reading fluency. At the bottom quartile of learning improvement, both control and intervention school learners could read half of a word less on average at endline. For the quartile of learners who improved the most in the respective groups, control school gainers improved by 22 words read correctly per minute, two less than the improvement for the fastest gainers in the treatment schools. On average, learners in intervention schools improved 1 word per minute more in reading fluency than the comparison group. The difference in the speed of learning gains for reading fluency is largest at the top 25 percent of the respective distributions of learning gains (approximately two correct words per minute).

A similar pattern holds for the distribution of improvements in reading comprehension scores between Grade 2 learners in the two groups. In both groups the bottom quartiles of learners in terms reading comprehension score improvements, saw their scores decrease by one correct answer over the year. For the two groups' respective top quartiles in terms of reading comprehension gains, intervention school gainers improved their comprehension scores by 8 correct answers over the year – one more than what the top gainers in control group managed.

Overall, based on the improvement in reading fluency and comprehension outcomes over the year, Grade 2 learners in intervention schools saw their reading proficiency outcomes improve at a faster rate on average than learners in control schools. The intervention school gainers' additional improvement over outcomes in control schools were largest amongst those learners in the intervention schools who improved the most. The largest differences between the two groups' respective gainers were in the upper half of the distributions of improvements.

7.2.4. HETEROGENOUS TREATMENT EFFECTS

This section investigates whether the intervention had any differential impacts based on learner-level characteristics, or more formally: whether there were any heterogeneous treatment effects. Results from other structured pedagogical programmes similar to the Funda Wande intervention (like the EGRS

34. The risk to investigating multiple possible sources of differential treatment effects is called data mining. In other words, if we test for differential treatment effects by a whole range of characteristics and sub-combinations of them, we increase the probability of finding a statistically significant result just by chance. However, if heterogenous treatment effects are found on some sub-group at both midline and endline assessments, we would be more confident that it is indeed a genuine effect. The analysis of heterogeneous treatment effects at this stage is thus limited to arguably the two most pertinent and policy-relevant learner level characteristics in this context.

35. A scenario where the two methods would lead to divergent results is in the case where the treatment has large heterogeneous treatment impacts based on learner's baseline reading proficiency, leading to a lot of rank mobility, but without moving the overall distribution of midline reading proficiency. This would be the case, for example, if learners changed rank due to treatment impacts (with originally weaker learners moving up the distribution, and originally stronger learners moving down the distribution), but the shape of the overall distribution of reading proficiency still looked the same at the end.

36. More specifically for the composite reading proficiency score: the greatest difference between learners in the two groups are between the 56th and 62nd percentiles, where intervention group learners consistently score 0.3 s.d. or more than their counterparts in the control group.

Table 10: Gains in reading comprehension by treatment status (Grade 2 only)

Quartiles of gains	Control gain	Treatment gain	T-C
Bottom 25%	-0,9 (1,5)	-1,0 (1,7)	-0,1
26th -50th percentiles	1,0 (0,0)	1,6 (0,5)	0,6
51st-75th percentiles	2,8 (0,8)	4,0 (0,8)	1,2
Top 25%	7,2 (1,5)	8,3 (1,8)	1,1
Total	2,1 (3,5)	2,5 (4,1)	0,4

and RUCS studies in South Africa and the Tusome studies in Kenya), and educational interventions more broadly, often find differential impacts on certain sub-groups. Two general themes arise from the literature. First, many programmes have the greatest impact on the already better performing learners, those who are better equipped to take advantage of the programme (Cilliers et al., 2019, Fleisch et al., 2017). Alternatively, some programmes seem to have the greatest impact on the weakest learners, those who often lag behind curriculum prescribed levels of learning and still need development in certain foundational skills. Second, in the South African context girls tend to outperform boys across all Grades, with the divergence in outcomes between the two groups evident right at the start of their schooling careers. As an illustrative example of the relevance here, the authors of the EGRS I study in South Africa collected contextual information on a range of learner, school and community characteristics for which they investigated differential treatment impacts (Taylor et al., 2017: 86-108). Of these, two key findings were that i) the effect of the programme was the greatest in the middle and upper parts of the distribution of baseline learner proficiency (with no significant effects found for the weakest learners) and ii) suggestive evidence that the structured pedagogy programme may be helping boys narrow the gap between them and girls. Given these considerations, as well as the data and sample size available to do a heterogeneity analysis this context, two learner level characteristics are investigated for any differential treatment effects at this stage of the evaluation[34]: i) baseline reading proficiency and ii) gender.

7.2.4.1. BY BASELINE READING PROFICIENCY

There is more than one way to see whether intervention impacts differ across the distribution of reading proficiency. One method would be to compare the entire midline distributions of literacy outcomes for the intervention and control groups, to see whether the "gap" in outcomes vary substantially at different points across the distribution. The second would be to check whether the impact of the programme varied with learners' baseline level of reading proficiency. The two approaches should yield similar results to the extent that treatment impacts are non-negative and fairly consistent across the distribution of baseline reading proficiency[35].

Looking back to ***Figure 23*** and

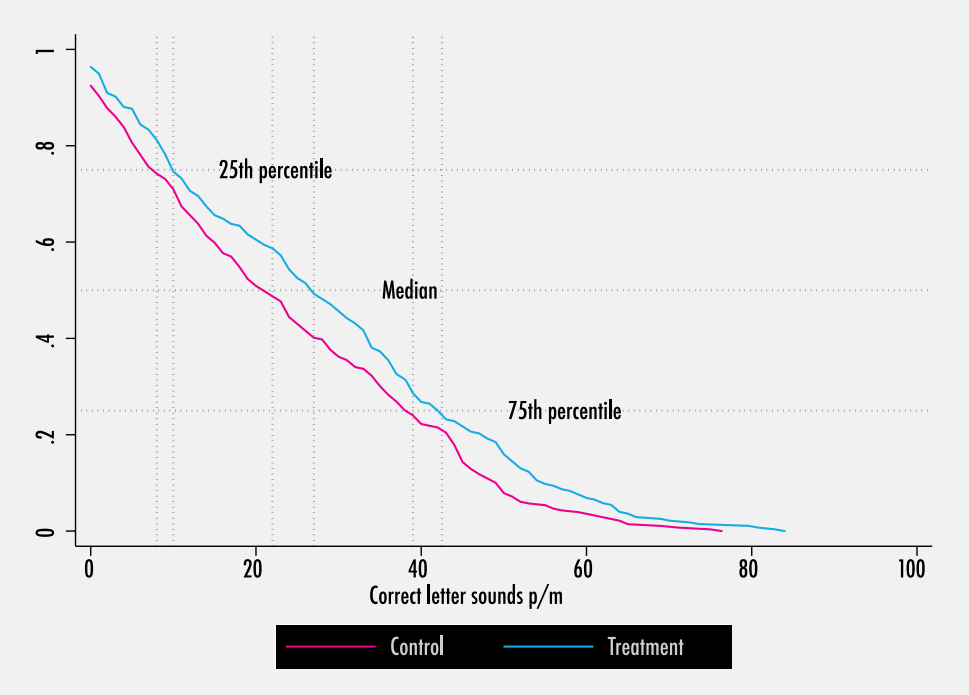


Figure 32: Correct letter sounds per minute by intervention group – Grade 1

comparing the distribution of midline scores for learners in the intervention group to their counterfactual, the control group, suggests that the programme impacts were positive across the distribution of reading proficiency. The largest impact seems to be concentrated in the mid-range of the distribution[36]. ***Figures 32 and 33*** show the percentage of learners scoring at or below a certain level for selected Grade relevant reading proficiency subtasks by intervention group status.

More specifically, ***Figure 32*** indicates that a greater share of intervention school Grade 1's could identify a certain number of correct letter sounds per minute at every point along the distribution. At the bottom of the distribution, 93 percent of control group Grade 1's could identify more than one letter sound correctly, compared to 97 percent of learners in the intervention group. At the higher end of the distribution, 42 percent of intervention school Grade 1's could read 34 correct letter sounds per minute or more, compared to only 34 percent of Grade 1's in control schools[37].

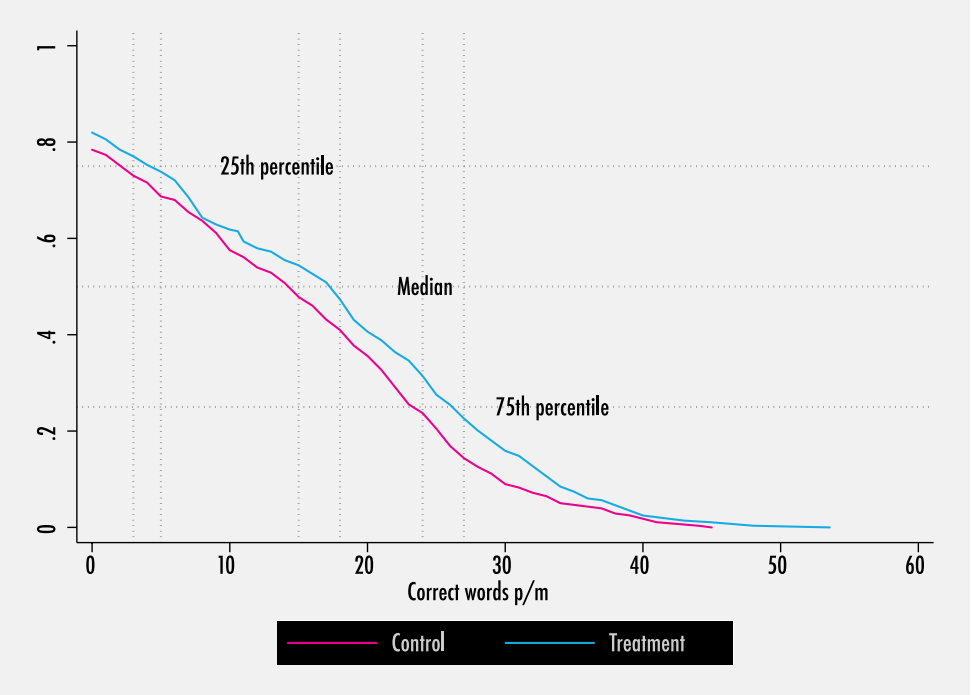
Another way to read the graph is to compare the correct letter sounds per minute for learners at the same position in their respective group's distribution. At the 25th percentile, intervention school Grade 1 learners could read 2 correct letter sounds per minute more (ten versus eight). At the medians there is a five letter sounds difference (27 versus 22), with a three-and-a-half letter sound difference at the 75th percentiles (42.5 versus 39). The largest consistent difference between the groups was in the middle of the distribution (a seven letter sound difference from the 55th to the 59th percentiles in the two groups). ***Figure 33*** performs the same exercise, but for the distribution of correct familiar words read per minute by Grade 2 learners in the two groups. Again, learners in the intervention group outperform control school Grade 2's across the entire distribution. At the lower end of the distribution, 81 percent of learners in treatment schools could identify at least one word correctly per minute at midline, compared to a slightly lower 78 percent of the control group's Grade 2s. At the 25th percentile of the respective distributions, intervention school learners read two words more correct words per minute (five, as opposed to three in the control group). At both the medians and the 75th percentiles, intervention school learners read

37. There is an emergent literature which suggests that the range around of 34 correct letter sounds per minute is the level of letter sound recognition characteristic of emergent- and basic readers in isiZulu (Spaull, Pretorius and Mohohlwane, 2020:12), a language which is very similar in structure to isiXhosa.

38. The baseline reading proficiency measure is a composite score constructed in the same fashion as the midline composite reading proficiency index (i.e. the first Principal Component from a Principal Component Analysis). However, it is constructed from only those tasks evaluated on both Grade 1 and Grade 2 learners at baseline. It also excludes i) any tasks that have severe floor or ceiling effects, or ii) any tasks for which an exploratory factor analysis indicated that the variable had a very low correlation with the other subtasks and loaded higher on the second underlying factor that seems to be indicative of oral literacy skills (and not of reading proficiency). For a detailed outline of the process, see pages 28 to 34 of the baseline report (Ardington, 2019) The baseline composite score was subsequently constructed from four common reading proficiency tasks at baseline: correct letter sounds per minute, correct digraphs and trigraphs per minute, phonemic awareness and a learner's ability to write letters.

39. The model in column (2) of Table 7 therefore also includes one additional control variable: the squared score of learners' baseline reading proficiency score, alongside the additional interaction term of intervention status with the squared baseline reading proficiency score.

Figure 33: Correct familiar words per minute by intervention group



three words more correctly per minute (18 versus 15 at the medians, and 27 versus 24 at the 75th percentiles). The largest differences are concentrated around the the mid-to-high range of the distribution.

To investigate whether the impact of the programme is statistically significantly different depending on learner's baseline reading proficiency[38], we rerun the same models as in the main analysis, but now adding interaction terms for treatment status and the learner's baseline level of reading proficiency. The same set of controls are used, including a separate control for each relevant measure of reading proficiency collected at baseline. ***Table 11*** below reports the coefficients on the school's intervention status, the interaction terms testing for heterogeneous effects based on baseline reading proficiency, the p-value of the test of whether the latter is statistically significant, as well as the sample on which the analysis was conducted.

Column (1) in ***Table 11*** tests whether there is a linear relationship between the intervention effect size and baseline reading proficiency. This suggests that there is no linear relationship between learners' baseline reading proficiency and the effect of the programme, given that the interaction term is not statistically different from zero (p-value=0.344). Column (2) tests for a quadratic relationship between the

Table 11: Intervention effects by baseline reading proficiency

	Outcome variable: Midline Composite score	
Treatment	0.174*** (0.054)	0.181*** (0.065)
Treatment x baseline composite score	-0.042 (0.044)	-0.032 (0.051)
Treatment x baseline composite score squared		-0.007 (0.026)
Sample	FULL	FULL
Observations	1,104	1,104
R-squared	0.678	0.684
Heterogeneous treatment effect: P-value	0.344	0.593

intervention and learners baseline reading proficiency[39]. It might well be the case, for example, that the programme has no effect on the weakest or best performing learners at baseline, but that the impact of the programme increases and peaks as one moves to the middle of the baseline reading proficiency distribution. However, a test for the joint significance of the two interaction terms in column two suggests that this is also not the case (p-value=0.593). Overall, the results suggest that the impact of the programme is positive and consistent across the distribution of learners' baseline reading proficiency.

We also graphically investigate the relationship between the programme effect size and the percentile rank of students based on their baseline reading proficiency in ***Figure 34*** below. The graph displays the estimated programme impact at each percentile of baseline reading proficiency[40]. The graph reaffirms the results from the regression analysis: programme effects are positive and fairly constant across the distribution of learners' relative rank of baseline reading proficiency. If anything, the programme might have a slightly larger impact on learners in the mid-range of the distribution of baseline reading proficiency. This result is consistent with both the main findings from earlier comparisons of the distributions of midline reading proficiency scores between the two groups.

Whilst ***Figure 34*** displays the programme effects for learners at different points in distribution of baseline reading proficiency in the full sample, one might also be interested in whether the programme's effects are as consistent based on students' relative ranks within the same classrooms. Even though this is unlikely to occur in the scenario where programme impacts are consistent across the sample distribution of baseline reading proficiency (as is the case here), the programme could hypothetically still have the greatest effect for learners who tend to lie at a certain end of the spectrum within their class. For example, intervention teachers might be both more effective in general and devote their attention disproportionately to helping students lagging behind to catch up. If what counts as "lagging behind" varies significantly across different classrooms and schools, and a large share of learners clump at the

40. More precisely, Figure 14 displays the local polynomial regression estimates of the effect size at each percentile of baseline reading proficiency. The estimates are obtained by first creating a value-added measure of reading proficiency, constructed by subtracting each learner's predicted score (based on the range of baseline covariates included in the estimation model) from their actual midline reading proficiency score. The value-added measure of the intervention is therefore equal to the difference (the residual), which is assumed to be attributable to the learner's intervention-status and other learner-level idiosyncrasy. Second, we estimate a local polynomial regression of the value-added measure on the percentile rank of baseline reading proficiency, separately for learners in the intervention- and control groups. The intervention impact estimate at each point in the distribution is obtained by subtracting the fitted values of each respective control percentile from the corresponding intervention percentile of student baseline reading proficiency. Finally, a pointwise 95 percent confidence interval is created using a bootstrap resampling of baseline percentiles (500 iterations), stratifying by sub-districts and clustering at the school level.

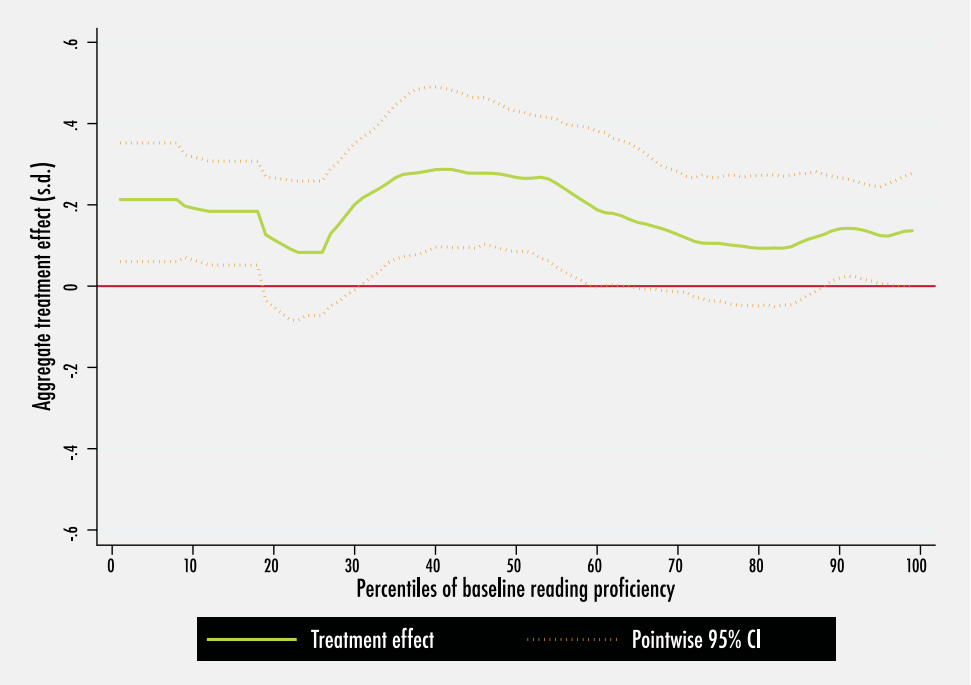
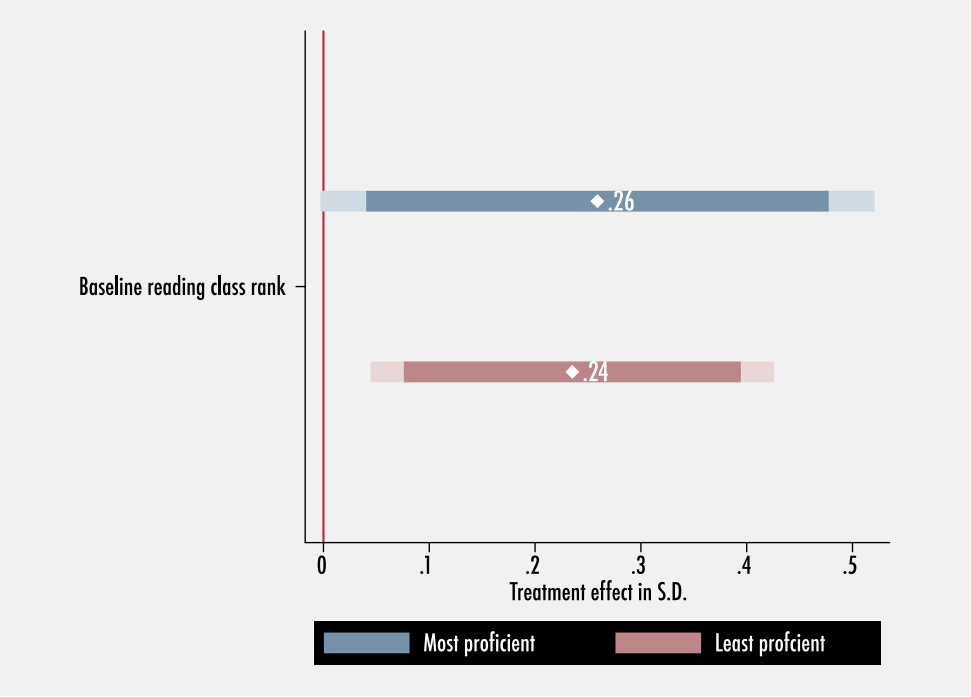


Figure 34: Nonparametric intervention impacts by baseline reading proficiency

Figure 35: Intervention Impacts estimated on most- and least proficient readers at baseline, respectively



bottom of their classroom's specific distribution at any one time, we could still see gains across the sample but which are concentrated at the lower end for specific classrooms.

Figure 35 below displays the estimated impact of the programme from two separate regressions for only those learners who ranked either first or last in their class for reading proficiency at baseline, respectively[40]. The intervention was clearly effective in shifting the reading outcomes of those learners who were the least proficient in their class at baseline (effect size of 0.24 s.d.). The estimated effect of the programme for the most proficient readers in a class at baseline is approximately as large (0.26 s.d.), but is more noisily estimated. Given the overlapping confidence intervals, we cannot conclude that the programme had any differential impact for a classroom's most or least proficient baseline readers at baseline.

Figure 36 repeats the same non-parametric estimation of treatment impacts based on learners' baseline reading proficiency as in ***figure 34*** above, but now based on a learner's baseline rank within the ten learners assessed per class. Treatment effects are consistent and positive, independent of where a learner ranked within their classroom at baseline. Together, these results suggest that neither students' absolute levels of baseline reading proficiency, nor whether they ranked at the top or the bottom of the class before the programme started, served as a constraint to the programme's effectiveness.

7.2.4.2. GENDER

A comparison of the distribution of midline reading proficiency scores for the intervention and control groups is shown separately by gender in ***Figure 37.*** Boys in both the intervention and control groups still score significantly lower on reading proficiency at midline than their girl counterparts in both control and intervention schools. ***Figure 37*** suggests that boys in the intervention schools do seem to be catching up with their girl counterparts, especially those boys in the upper half of the reading distribution for boys.

We test whether these apparent differential impacts are statistically significant by including an interaction term for intervention group status and whether or not a learner is female in regressions similar to those in ***Table 11. Table 12***, column one below shows that while the point estimate of the

41. The ranking methodology worked as follows. For the teacher identified "most proficient reader" (or top ranked learner), the learner(s) who scored the highest mark in the class on the relevant task were assigned a ranking score of one. If learners were tied for first place, both would receive a score of one. All subsequent learners would receive a ranking score of one plus the number of learners between them and first place. In the example where two learners were tied first, the next-best ranking learner would receive a rank score of three, and so on. The ranking methodology for the teacher identified "least proficient reader" (or bottom ranked learner) worked in exactly the same manner, but with the learner with lowest score in the class now receiving a rank of one. If, for example, eight learners in the class all scored zero on a task, they would all receive a rank of one, with the next worst learner receiving a rank score of nine. Learners with tied scores at the top and (especially) at the bottom of the class presented challenges to the construction of rank scores based on tasks like learners' correct letter sounds identified or reading fluency. The Grade 1 class rank used to determine the most and least proficient readers follows the same ranking methodology as outline in section 7.3.1 below, but based on baseline reading proficiency scores. Baseline reading proficiency scores are derived using PCA of letter sound recognition-, phonemic awareness and copying letters tasks (i.e. the same index score as the baseline composite score used in table 7). The Grade 2 rank is based on a reading proficiency score constructed similarly, but including word recognition-, reading fluency- and comprehension tasks from baseline assessments for Grade 2 learners specifically. This means the Grade 2 reading proficiency rank is based on similar sub-domains of reading proficiency as the midline composite reading proficiency score. Excluding these baseline tasks from the Grade 2 composite score would have limited our ability to differentiate the most proficient reader among those Grade 2 learners in a class who were already highly proficient in lower order decoding tasks, but might vary on how they perform in higher order word recognition, paragraph reading and comprehension tasks. In classrooms where more than one learner was tied as the least proficient reader, all the least proficient readers in the class are used in the analysis but down-weighted in proportion to how many learners a classroom is contributing to the analysis (i.e. inverse probability weighting).

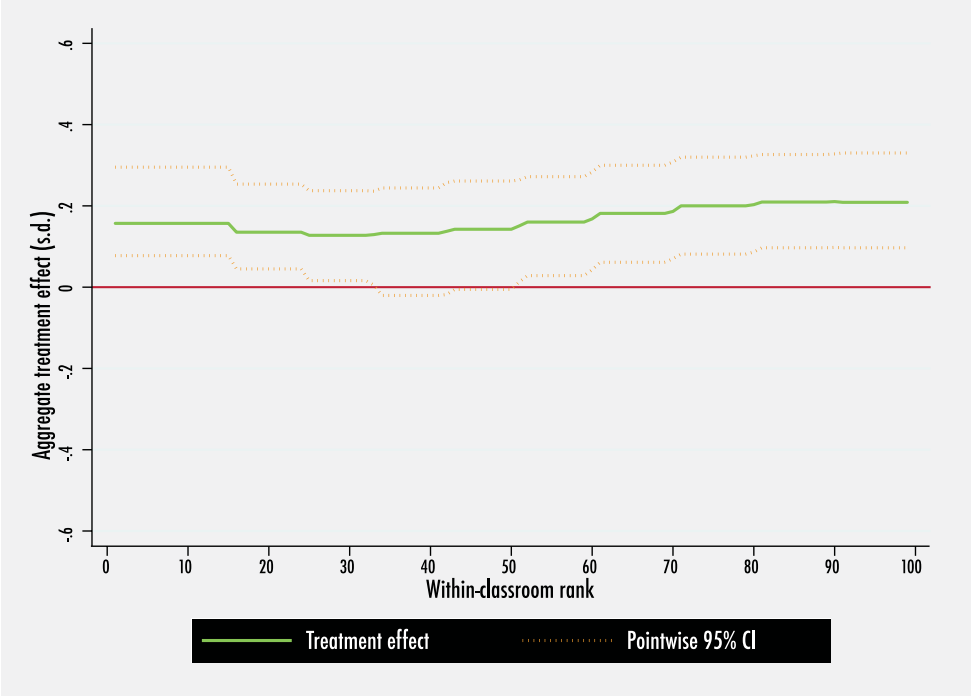


Figure 36: Nonparametric intervention impacts by within class rank of baseline reading proficiency

treatment effect is higher for boys than girls in the full sample (0.20 s.d. versus 0.15 s.d.), the interaction term does not yield a significant estimate - suggesting that there is no differential impact between boys and girls on average. However, we find stronger evidence of differential treatment impacts by gender when the analysis is redone on each Grade separately. More precisely, columntwo in ***Table 12*** shows that the point estimate for boys in Grade 2 is 0.27 s.d., whilst the impact on Grade 2 girls' reading proficiency is only approximately 0.08 s.d.. This difference in intervention effects by gender is significant at the 90 percent confidence level. The inverse is the case for the point estimates on the Grade 1 sample, where boys have a point estimate of the treatment impact of 0.15 s.d., 0.13 s.d. less than the impact estimate for Grade 1 girls. This differential intervention impact is not significant, however.

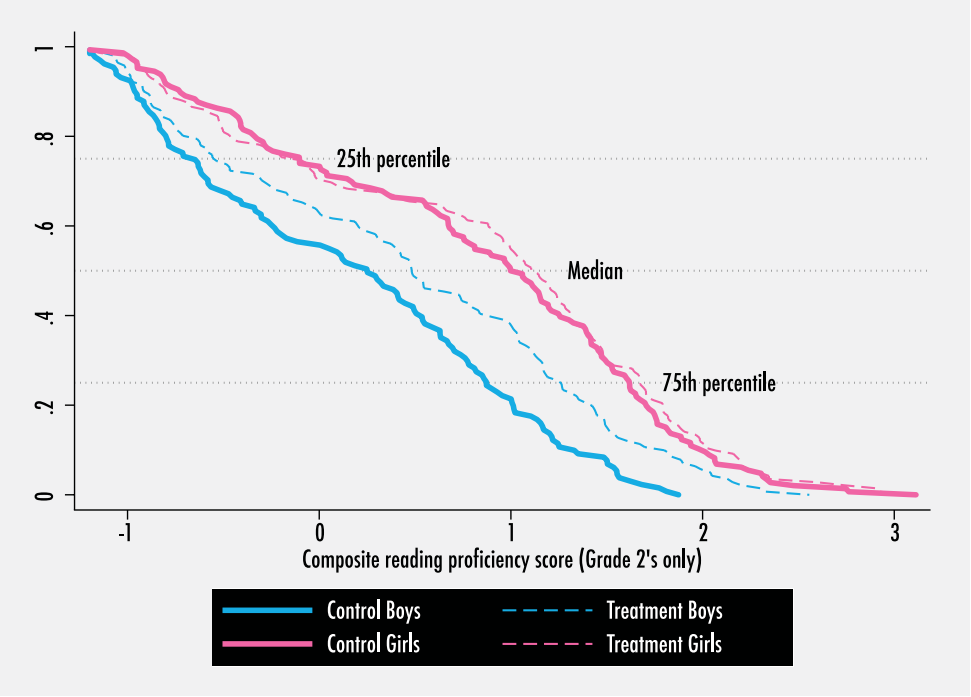
Altogether, these results would suggest that the programme has the greatest impact on learners at a certain stage of their literacy development trajectories. For example, Grade 1 girls have a higher baseline level of reading proficiency on average, and more of them might therefore be at a certain point in their developmental paths at which higher order literacy skills (like reading fluency) have the potential to be acquired within a short time frame, giving the right educational inputs. Boys, in turn, who are generally lagging behind their female counterparts, are perhaps more likely to only reach such a point where they have the potential for rapid skill acquisition a bit later, in Grade 2.

Digging deeper into the differential treatment effects for boys and girls in Grade 2, ***Figure 38*** suggests that boys at higher levels of baseline reading fluency experience much larger gains in reading fluency (over control group, boy counterparts) than what is the case for Grade 2 girls in intervention schools. The locally weighted polynomial regression shows that Grade 2

Table 12: Intervention effects by sex

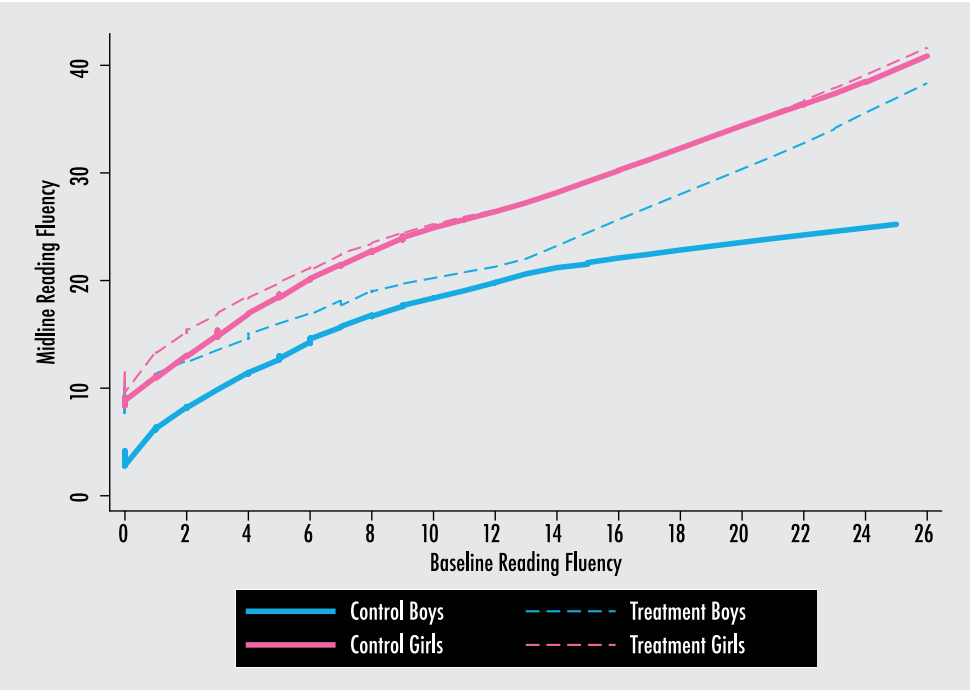
	Full	Gr 1	Gr 2
	Outcome variable: Midline Composite score		
Treatment	0.199*** (0.066)	0.269*** (0.069)	0.145 (0.090)
Treatment x female	-0.050 (0.076)	-0.194* (0.098)	0.128 (0.094)
Sample	FULL	Grade 2	Grade 1
Observations	1,104	552	552
R-squared	0.678	0.756	0.428
Heterogeneous treatment effect: P-value	0.507	0.054	0.178

Figure 37. Composite reading proficiency by intervention status and sex (Grade 2)



intervention school girls at higher levels of baseline reading fluency see basically no gains in reading fluency over their control school girl counterparts who were at similar levels of baseline fluency. In contrast, Grade 2 boys in the intervention group have large gains over the comparison group boys at similar levels of baseline reading fluency, both at the bottom-most levels of baseline reading fluency (less the five correct words per minute), and to an even greater (and increasing) extent at higher levels of baseline fluency (from approximately 14 correct words per minute and up at baseline). This suggests that Grade 2 boys in intervention schools were able to narrow the gap with Grade 2 girls in intervention schools with similar levels of baseline reading fluency, especially if the boys started with reasonably high levels of baseline reading fluency.

Figure 38: Greater gains in reading fluency for best performing boys at baseline (Grade 2)



7.3. Intermediate Outcomes: Teachers

Pedagogy in the majority of South African classrooms is largely communalized. Teachers tend to work "with the whole class as a homogeneous group, with little or no differentiation of tasks or differentiation of individual performances. In a communalizing pedagogy, close scrutiny of what individual learners produce is not possible because productions are generally chorused, and, consequently, individual assessments of learners' progress in reading is foreclosed" (Hoadley 2019). In this section we explore whether one of the pathways through which the intervention may be achieving impacts on learners' reading skills is through better formative assessment of learners' ongoing progress in intervention schools. We also examine whether there are any shifts towards more individualised forms of learning through teachers reports on their use of the Vula Bula Graded reader anthologies.

7.3.1. FORMATIVE ASSESSMENT OF LEARNER PROGRESS IN READING

Teachers were asked to identify both the most proficient and least proficient readers from a list of the approximately ten randomly selected learners that were assessed in their class. For each learner that a teachers ranked as the most/least proficient reader, we can therefore compare where they actually ranked out of the learners assessed based on an objective measure of reading proficiency from the learner assessments. From here on in, the teacher identified learners will be referred to as the "most proficient readers" and "least proficient readers", respectively. The ability to correctly identify the those learners who are truly the most/least proficient readers out of the ten learners assessed per class can then also be compared between teachers in the control and treatment schools. An important caveat is that the sample is not powered to detect relatively small differences in the actual rankings of "most proficient-" and "least proficient readers" in the two groups, given that only 108 teacher observations (and thus only so many "most proficient" and "least proficient" estimates) are available[42].

In terms of measurement, no exact definition or criteria on how to determine the most/least proficient readers in the list was provided to the teachers. An obvious candidate would be oral reading fluency. However, 55 percent of Grade 1's[43] scored zero on the reading fluency task and we assume that this may have complicated rankings (and especially identifying a least proficient reader) for teachers of Grade 1 classes, where the majority of learners could not read a single word from a passage. We therefore use the midline composite reading proficiency measure as this better discriminates at the bottom of the distribution. We used learners' actual rank based on this composite measure of proficiency for the comparison of "most proficient" and "least proficient" identified readers actually performed[44].

When comparing the accuracy of the rank between teachers in control and treatment schools, teachers in the intervention schools

42. Three teachers said that they did not know who the most proficient readers in the list of ten is, whilst one teacher also indicated that they did not know who the least proficient reader is. Furthermore, one learner who a teacher estimated to be the most proficient reader was absent on the day of midline assessments, and one teacher identified "least proficient reader" could not be assessed at midline due to behavioural and/or learning disabilities which prevented the learner assessment from taking place. Data is thus available for 104 observations on the teacher identified "most proficient readers", and 106 observations on the teacher estimated "least proficient readers" variable.

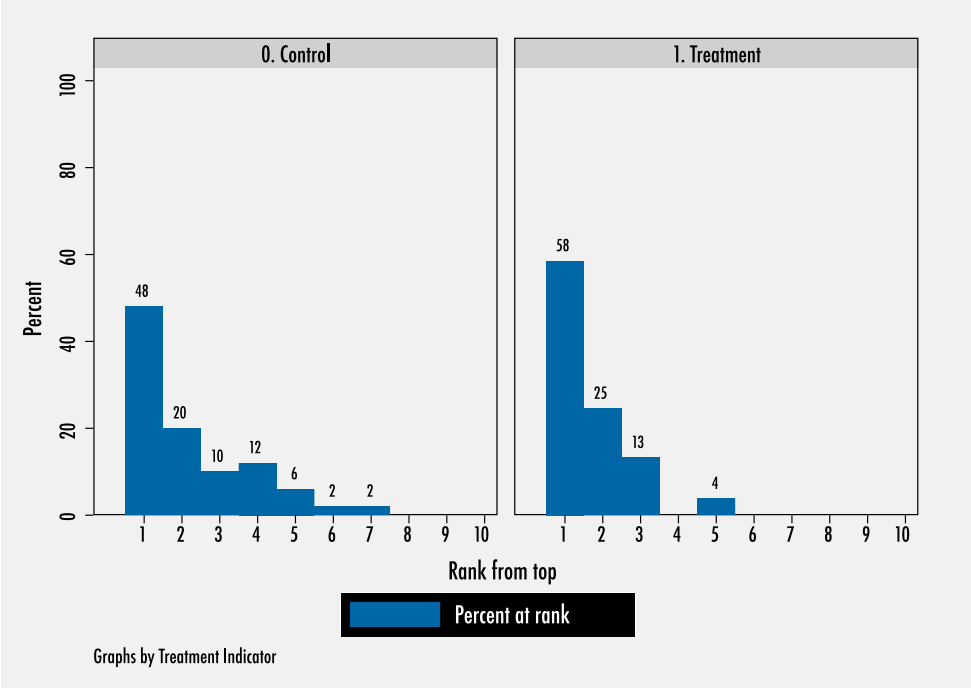
43. In comparison, only 17 percent of Grade 2's scored zero on the Oral Reading Fluency task.

44. The class rank used to determine the most and least proficient readers follows the same ranking methodology as outline in section 7.2.4.1 above, with class ranks constructed based on the composite reading proficiency measure at baseline. There were no ties at the top and only one classroom where three learners were tied for bottom for reading proficiency scores.

45. The regression included strata fixed effects and standard errors were clustered at the level of the school.

46. We conducted similar analyses using correct letter sounds per minute for Grade 1 learners and oral reading fluency for Grade 2 learners. There was no significant difference in the ability of treatment and control teachers to identify the weakest learners using either of these measures.

Figure 39: Rank of teacher identified "most proficient readers"



were more accurate on average in their predictions of the most proficient reader. ***Figure 39*** shows the distribution of where "most proficient readers" actually placed for reading proficiency within the 10 assessed learners in their class. Fifty-eight percent of treatment teachers identified the top ranking learner (rank of 1) and a further 25 percent selected the second ranked learner. The comparable figures for control teachers are 48 and 20 percent. We formally tested whether treatment teachers were better able to identify the most proficient readers in their class using an ordered logit regression[45] of the rank of the learner that the teacher identified as most proficient on an indicator that the teacher was in an intervention school. Treatment teachers are significantly more likely to identify a learner with a better rank as the most proficient learner (p-value = 0.05) (output shown in appendix ***Table A5)***[46].

Next we examine teachers' ability to identify the "least proficient reader" of those assessed in their class. Now learners' actual rank is defined in terms of how far they rank from the bottom of the distribution (i.e. a rank of one implies the learner had the lowest score on the composite measure). From ***Figure 40***, a greater proportion of the control school teacher identified "least proficient readers" are indeed the bottom ranking learner (i.e. those with a rank of first from bottom - 47 percent), compared to treatment teacher identified "least proficient readers" (36 percent). However, there is much more variability in the actual ranks of the learners selected by control teachers than those selected by treatment teachers. Eighty-five (85) percent of treatment teachers select the bottom three learners, as opposed to 78 percent of control teachers. As with the "most proficient readers", we conduct a formal test using an ordered logit. The point estimate is negative suggesting that treatment teachers are better able to identify poorer performing learners, but the estimate is not statistically significant (output shown in appendix ***Table A5***).

There are at least two, somewhat related reasons why teachers in intervention schools seem to have become better able to distinguish the top-performers among those assessed in their respective classrooms, whilst their relative ability to identify the least proficient readers is less clear. The first relates to the nature of the measurement and ranking methodology itself. The composite reading proficiency measure used to construct rankings is likely noisier in identifying the truly least proficient reader. Given the lack of variation in outcomes, the measure is relatively less able to accurately differentiate true proficiency among those learners that clump at the bottom of the distribution in classrooms where more than one learner has low-to-zero scores across a range of the reading proficiency sub-tasks.

Second, intervention teachers might have become relatively better in picking out the best among those few learners who perform towards the upper end of the class's reading proficiency distribution. There seems to be more variation in reading proficiency for teachers to observe among the sub-set of learners whose reading ability is towards the top of a class's distribution. In other words, it might be the case that it is inherently a much easier task to differentiate between the few highly proficient learners in a class, conditional on the teacher being able to conduct formative assessments and become aware of learners' differential reading abilities. In contrast, it might be an inherently more difficult task to differentiate among a larger share of learners who clump at very low levels of basic reading proficiency skills (i.e. the multiple learners who basically can't perform any observable reading tasks). In sum, intervention teachers' increased ability to differentiate among learners at different reading proficiency levels might therefore be an increasing function of how well the learner performs.

Teachers were also presented with

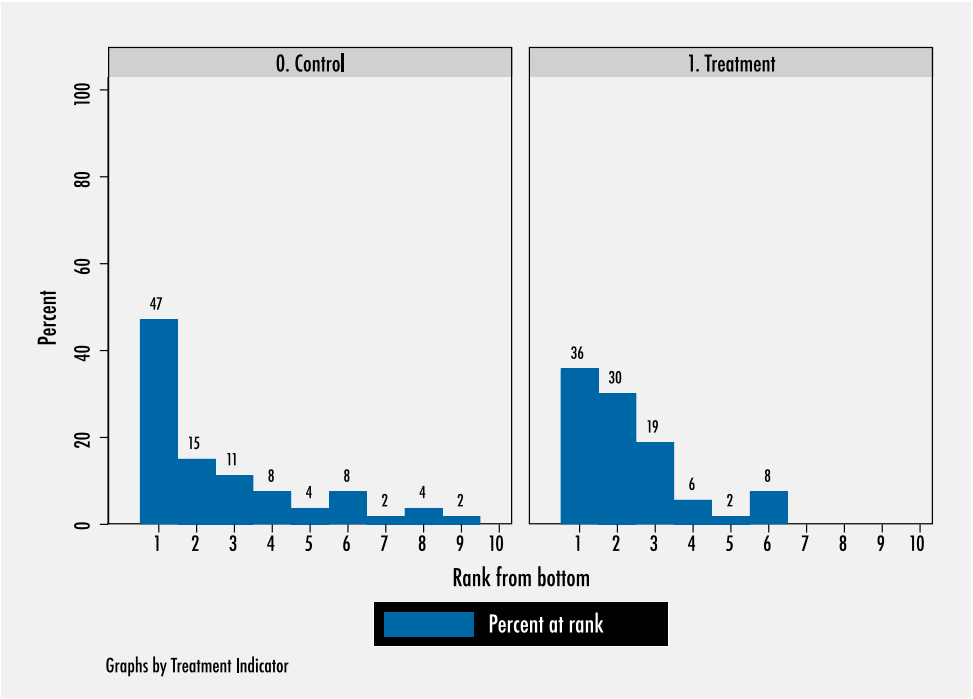


Figure 40: Rank of teacher identified "least proficient readers"

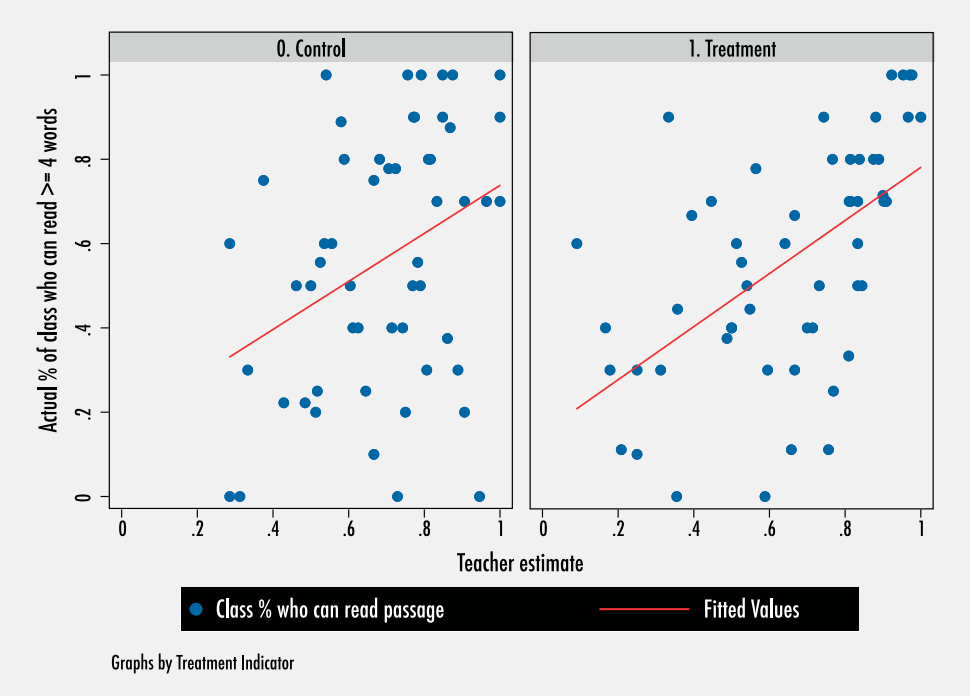
three passages from the Vula Bula anthologies, with a passage each from the Grade 1, Grade 2 and Grade 3 level anthologies. Based on the passage presented to them, teachers were asked how many learners in their class would be able to read the specific passage. Given then that we have information on each teacher's class size, we could translate their estimate into a percentage score of the share of the class who could read a specific passage. On average, Grade 1 teachers felt that 60 percent of their learners could read the Grade 1 text, but only 48 percent and 37 percent would be able to read the Grade 2 and Grade 3 texts respectively ***(Table 13)***. A similar percentage (64 percent) of Grade 2 teachers thought their learners could read the Grade 2 text. On average, Grade 2 teachers feel that around one quarter of their learners would not be able to read the Grade 1 level text on their own. Grade 2 teachers believe that one in two of their learners would manage to read the Grade 3 text on their own.

Given then that we have data for each teacher on 10 randomly selected learners from their class, we can compare how actual reading levels (i.e. the actual amount of words read correctly from the reading fluency task) compared with the percentage of learners the teacher believes can read the passage independently. The Grade 1 level passage requires learners to be able to read a least four words correctly, whilst the Grade 2 passage consists of 18 words. We construct two indicators of the share of the 10 assessed learners in each teachers' class that can read at least 4 and 18 words or more.

Figure 41 graphically presents how well teachers' estimates of the share of the class who can read the Grade 1 level passage (on the horizontal axis) compared to the share of learners assessed who could indeed read the passage (on the vertical axis), for classrooms in intervention and control schools separately. A line of best fit is drawn in each panel, obtained from a regression of how well teacher estimates predict the actual shares of the 10 randomly

Table 13: Percentage of learners in class able to read text at this level

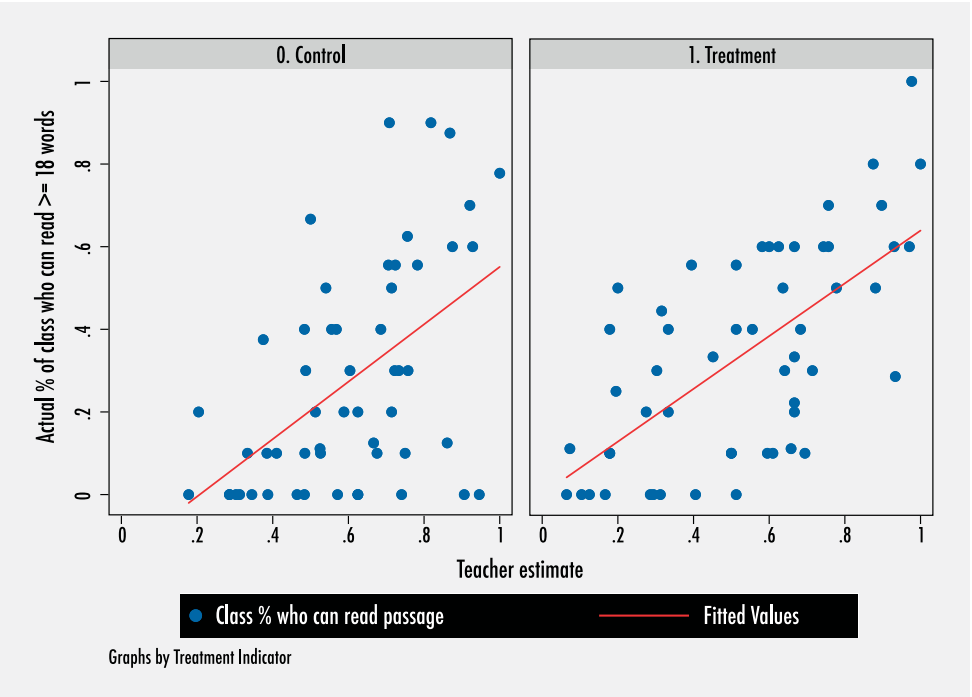
Figure 41: Percent of learner who can read >=4 words per minute over teacher report on percent who can read Grade 1 passage, by treatment status



selected learners in their classroom who could read the specific passage. The same exercise is repeated for teachers' estimates on the share of their class who can read the Grade 2 level passage (i.e. at least 18 words correctly) in ***Figure 42.***

For both passages, teachers in intervention schools are better at predicting the share of their class who can read at the level required to actually read the relevant passage. For the Grade 1 passage, the goodness-of-fit (or R-squared) value of teachers in intervention schools is 0.32, compared to 0.13 for teachers in control schools. In other words, intervention school teachers' predictions explain 32 percent of the variation in their learners actual ability to read the Grade 1 passage. For the Grade 2 level passage, teachers in both groups are better able to predict the share of their class who can read the passage, with R-squared values of 0.27 and 0.43 for control and intervention classrooms, respectively. The intervention school

Figure 42: Percent of learner who can read >=18 words per minute over teacher report on percent who can read Grade 2 passage, by treatment status



teachers were again significantly better predictors of their classrooms' actual reading abilities.

Despite intervention school teachers being the better predictors, the predictions in both groups were pretty inaccurate. This can be gleaned from the low R-squared values, but also simply from inspecting Figures 41 and 42. Classroom outcomes deviated significantly from the estimates from teachers on what share of their class could read a given passage. Also worth noting is the fact that teachers in the intervention group were more likely to report that almost none of the leaners in their class could read a specific passage. For example, from ***Figure 42*** for the Grade 2 level passage, multiple intervention school teachers reported that less than 20 percent of their class could read at this level (and correctly so), whilst almost no control school teacher estimated that less than 20 percent of their learners could read at the Grade 2 level (even though this was most certainly the case in many their classrooms).

7.3.2. USE OF VULA BULA ANTHOLOGIES

Baseline data highlighted the challenges with access to reading materials at both school and home in both intervention and control schools. Around half (53 percent) of schools had no library, a quarter of learners had no library or bookshelf in their class and 65 percent of learners have no books other than schoolbooks to read at home. Almost all classes had pre-existing readers but the quantities were woefully inadequate for the number of learners in the class. In most classes there were insufficient readers for activities such as group guided reading or for learners to progress through the year. In 2019 for the first time, the ECDoE distributed a Grade-level Vula Bula anthology of 22 Graded readers to every Grade 1 to 3 learner in all their primary schools. Details on the distribution and use of the anthologies from the teacher interviews, classroom observational checks, principal interviews and learner interviews are provided in Appendix V. In summary, distribution was almost universal, most learners received their own copy of the anthology and were allowed to take them home and reported usage of the anthologies at school and at home was high.

Of the teachers who received copies of the anthologies, all but one report using them in class. ***Figure 43*** shows the frequency of reported use in intervention and control classrooms. Overall, reported use is high in both intervention and control classrooms. However, the frequency of use is substantially higher in intervention classrooms. Over a third of intervention teachers report using the anthologies daily in contrast to only eight percent of control teachers. Intervention teachers also report using more of the stories in the anthology than control teachers ***(Figure 44).***

Learners were also asked who often the anthologies were used in the class ***(Figure 45)*** and how many of the stories that had read (either themselves or by an adult) ***(Figure 46)***. Learner reports of frequency of use and stories read tend to be lower than that of teachers in both intervention and control classrooms. This is suggestive of some desirability bias in the teacher responses. Nevertheless, the learner reports accord with those of the teachers in pointing to higher use of the anthologies in intervention classrooms, although the differences are more muted[47].

Table 14 shows the range of activities that the teachers conduct with the anthologies by intervention status. In almost half of both intervention and control classrooms, the teacher reads aloud from the anthology while learners listen. The other form of communalised learning, the whole class reading from the book at the same time, is much more common in control than intervention classrooms (30 percent versus 15 percent). Treatment teachers are much more likely to report using the anthologies for group

47. There are other indicators that suggest the anthologies are more well used in intervention schools – 1) there is more variation in the choice of learners' favourite story from the anthologies, 2) learners are nine percentage points (92 percent versus 83 percent) to report that they have read a specific story shown to them by the enumerator, and 3) conditional on having read the story, they are 11 percentage points more likely to answer the simple question about the story correctly (71 percent versus 60 percent).

Figure 43: Frequency of use of Vula Bula stories in class – teacher report

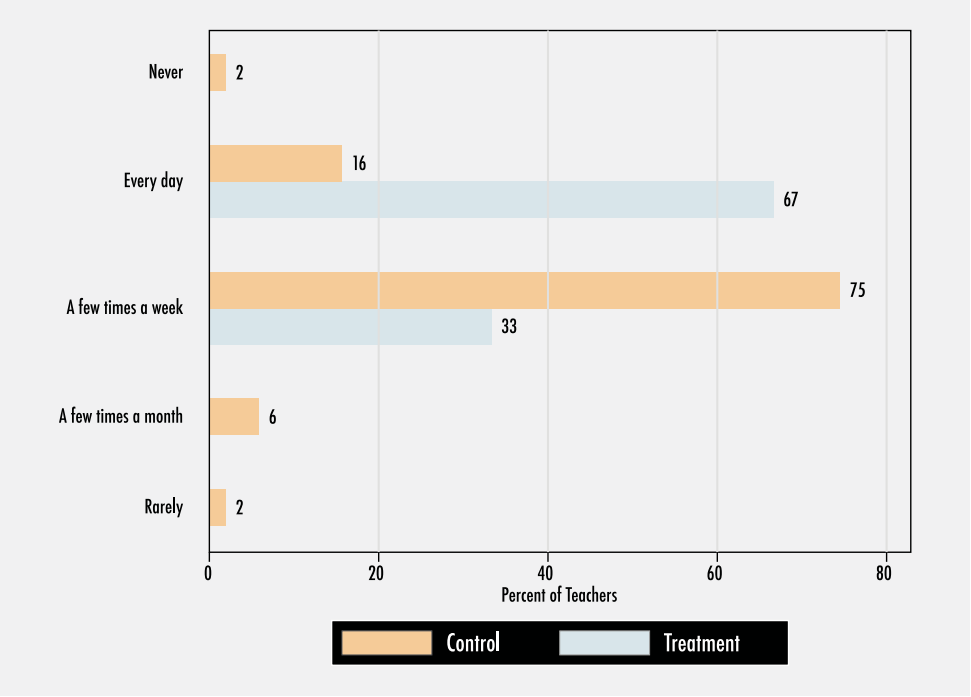


Figure 44: Number of Vula Bula stories used – teacher report

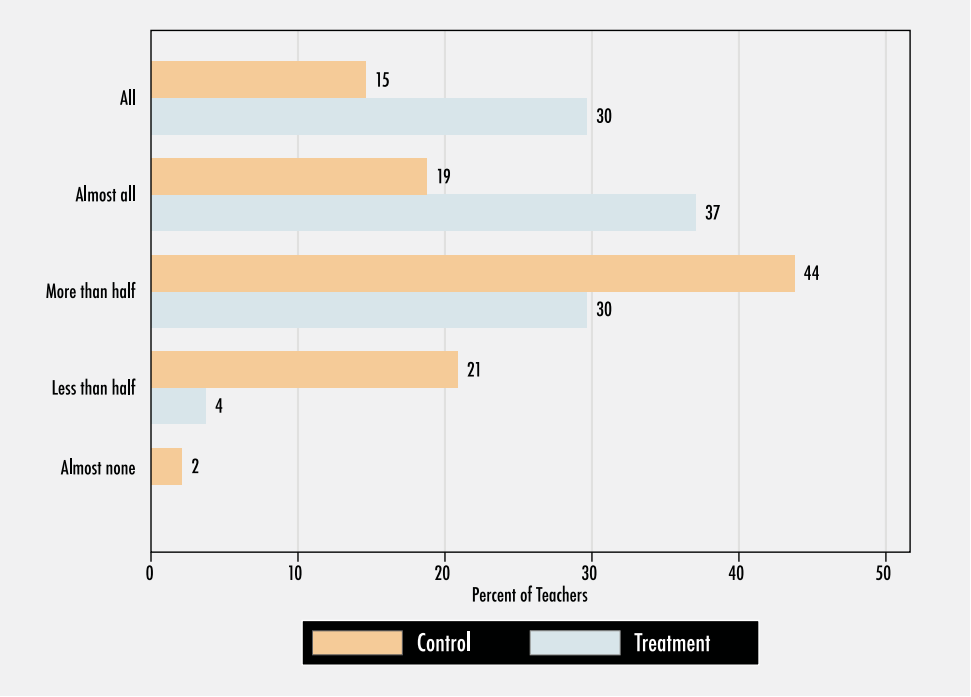


Table 14: Anthology use in class by treatment status – Teacher report

	Control	Treatment
Teacher reads aloud from anthology while learners listen	46%	46%
Whole class reads from book at same time	30%	15%
Small groups during group guided reading	52%	85%
Shared reading/paired reading	64%	72%
Independent reading during break/ DEAR period	58%	37%

guided reading in small groups (85 percent versus 52). They are also more likely to use the anthologies for shared and paired reading. It is possible that teachers in control schools use other reading resources for these activities. However, given the woefully inadequate quantity of readers available in both classrooms at baseline and the universal delivery and high reported general usage of Vula Bula, this is an unlikely

Figure 45: Frequency of use of Vula Bula stories in class – learner report

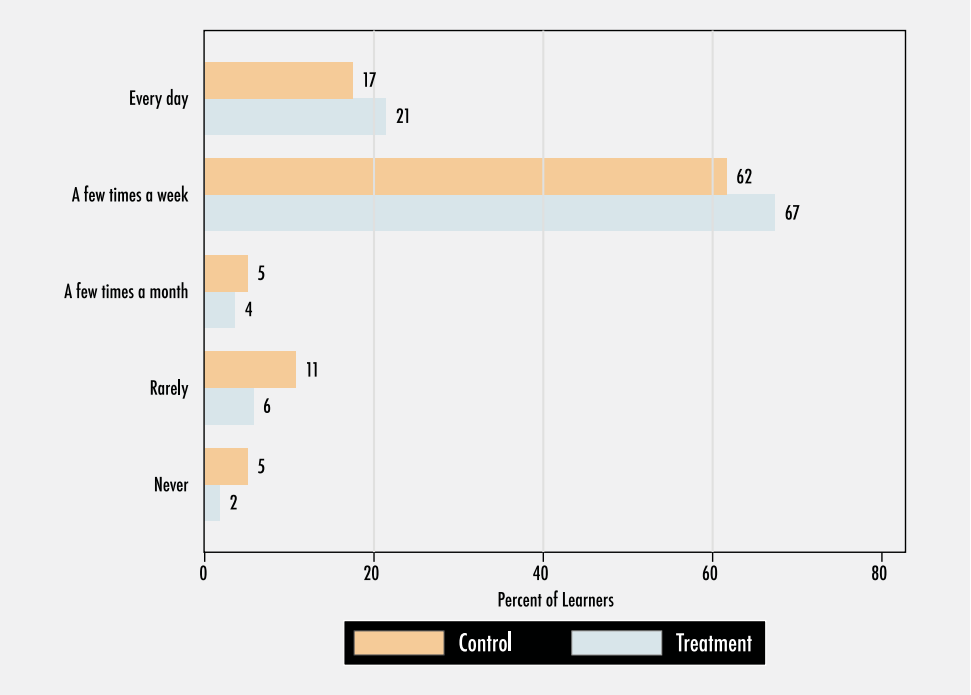
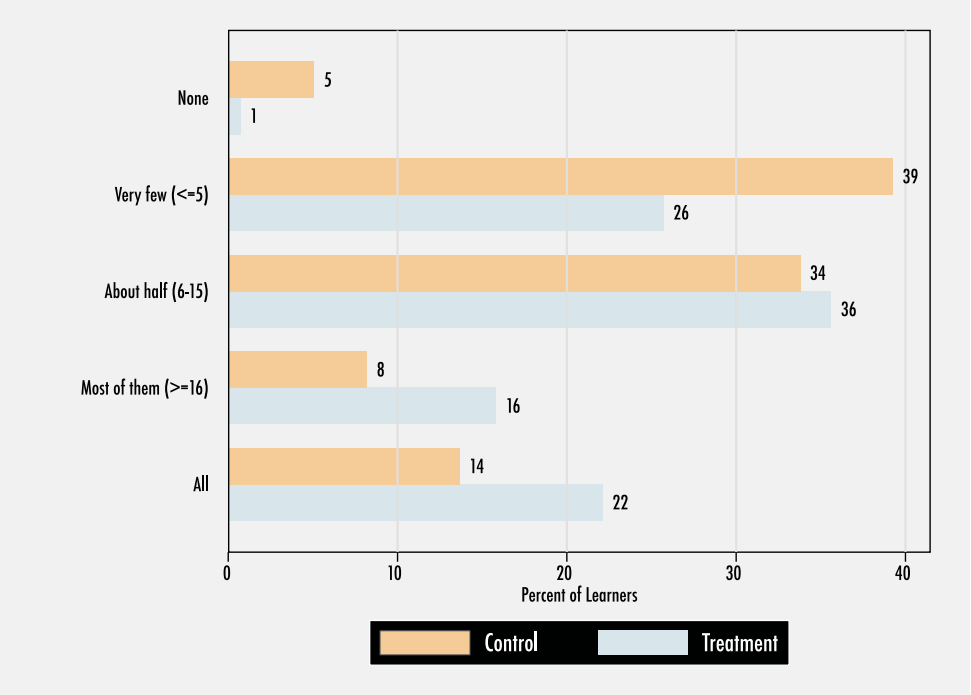


Figure 46: Number of Vula Bula stories used – learner report



explanation for the differences in use presented in ***Table 9.***

These use patterns seem to suggest a shift away from communalised learning towards more individualised modes of reading instruction. We do need to bear in mind that this information is self-reported and desirability bias (teachers reporting what they think they are expected to be doing rather than what they actually do) is plausibly higher among ***i***ntervention teachers. The *in-depth* qualitative classroom observations planned for term three in the second year of the intervention will shed more light on whether and how Funda Wande is shifting reading pedagogy in the classroom.

8. Discussion

8. Discussion

The Funda Wande intervention had a 0.17 s.d. impact on the learner's reading proficiency after one year of implementation. The programme impacts are positive across all the sub-tasks assessed, with impacts at this stage largest on certain foundational skills (like correctly identifying letter sounds and being able to manipulate phonemes); the skills that learners require to decode words, read more fluently and eventually read for meaning. Nevertheless, the impacts on higher order skills like word recognition and reading fluency are almost as large and significant.

In practical terms, learning gains on the subtasks on which the intervention had a positive effect translated to between 20 to 27 percent of a year's worth of learning for Grade 2 learners. For example, a two word per minute increase in familiar word recognition for Grade 2 learners in intervention schools is roughly equivalent to a school term's worth of learning in comparison schools. Learners in Grade 1 classrooms in intervention schools gained even more over their peers in comparison schools for the emergent- and pre-literacy skills on which the programme had positive effects. For letter recognition tasks, phonemic awareness and productive listening comprehension skills these Grade 1 learners' outcomes improved between 33 to 58 percent more than the 'business as usual' development in these skills that occurred in control schools. Concretely, intervention school Grade 1's could correctly identify six letters sounds per minute more after one year of exposure to the intervention, equal to a third of a year's worth of learning in control schools.

When investigated by Grade, certain dynamics of learners' learning trajectories in the different Grades and how these relate to the programme impacts come to the fore. For example, the intervention impacts on Grade 1 learners' foundational skills (letter sound recognition and phonemic awareness) are particularly large, both relative to the impacts on other Grade 1 literacy skills and the impacts on similar skills for Grade 2 learners. For Grade 2 learners, the impacts of the intervention are more consistent across foundational- (letter sound recognition) and higher order literacy skills (like word recognition, reading fluency and reading comprehension outcomes), but not significant on phonemic awareness. These dynamics suggest that certain foundational decoding skills, like letter-sound knowledge, phonemic awareness, and word recognition are important for learners to master before they can effectively progress toward reading passages fluently. The results support the idea that learners require a range of foundational literacy abilities before they can read with some level of speed and accuracy (i.e. fluency), and in turn, then need to read with a certain minimum level of fluency in order to comprehend what they are reading.

Consistent with the sequential nature of literacy skill acquisition, we only find positive and significant effects on reading comprehension for Grade 2 learners at this stage of the intervention's implementation. These estimates are, however, almost as large as the effect sizes on lower order literacy skills for Grade 2 learners, and

found consistently across more than one reading comprehension task. We would therefore expect to find similar effects on reading comprehension for 2019's Grade 1 learners at the next round of learner assessments at the end of two years of exposure to the programme.

At a practical measurement level, the evaluation has provided valuable lessons in how to appropriately measure reading for meaning. For example, the relatively extensive range of comprehension tasks used provide valuable insights into the extent to which the different levels of reading comprehension rely on reading fluency, whilst also illustrating the limited ability of current comprehension assessments to differentiate among those learners at very low levels of fluency. Certain tasks (like the vocabulary instrument) also proved less useful due to severe ceiling effects and thus an inability to differentiate among learners. Some tasks will therefore be updated and/or replaced in future rounds of assessment, also taking into account the relevance of tasks to learners in different levels of progression. This project will also feed into and build on existing empirical research on African languages in South Africa, with the longitudinal rounds of data on reading skill acquisition allowing a better understanding of the minimum level of decoding- and reading fluency skills required to progress to reading for meaning in isiXhosa.

Also encouraging from a policy perspective is that the intervention seems to have fairly consistent positive impacts for learners across the distribution of baseline reading proficiency. If anything, the intervention has slightly larger impact for those who are mid-range in terms of baseline reading proficiency. When looking at treatment impact by gender, the extent of the differential impact seems to rely on the Grade that the learners are in. Whereas the programme clearly helps boys in intervention schools catch up to their girl counterparts in intervention schools during the course of Grade 2, it seems that it might have the opposite effect in Grade 1 (helping intervention school girls pull further away from boys in the same schools).

With respect to the mechanisms at play, at this stage we find suggestive evidence across a multitude of indicators that teachers in intervention schools are more likely to a) be more attuned to the actual reading proficiency levels of the learners in their class (both in terms of whether learners are at the top or the bottom of the distribution and how the class performs overall) and b) to make greater use of material resources provided, and (c) to use instructional techniques that have previously been shown to facilitate more individualised forms of learner reading practice, feedback and learning (Cilliers et al., 2019).

For example, intervention school teachers clearly have a better sense of who the most proficient readers in their classrooms are. They also do better in estimating what share of their class can read a passage of a given difficulty level. With respect to resource use, intervention teachers seem to use the Vula Bula Graded readers more often and cover more stories. Finally, in terms of teachers' instructional practices, control school teachers are more likely to make use of certain communalised forms of learning, like having the whole class simultaneously read aloud from the same passage. In contrast, intervention teachers are more likely to report using the anthologies provided to practice group guided reading. This technique has been proven effective in allowing teachers to stream learners into groups by reading ability, differentially target the texts to a group's ability and provide readers with more individualised opportunities to read and receive feedback (Cilliers et al., 2019).

The midline results from the EGRS 1 study (see Taylor et al., 2017: 41-47) provides i) a second data point from which to start identifying common patterns and insights from structured pedagogic interventions in low resource South African schools,

A common theme in both the studies is that these structured pedagogic programmes first shift foundational decoding skills for Grade 1 learners, before relatively greater improvements on higher order domains of reading proficiency follow in Grade 2.

as well as ii) a point of reference against which to compare the Funda Wande programme effectiveness after one year of implementation. After roughly four terms of programme implementation, the estimated effect of the Funda Wande intervention is at least as large as that of the EGRS 1 intervention's most effective treatment arm, the coaching intervention (after a somewhat shorter three terms of programme implementation)[48]. It is important to note, however, that in both cases the confidence intervals around the estimated programme impacts at midline overlap and one cannot say with any level of certainty that the effect of one programme was larger.

A common theme in both the studies is that these structured pedagogic programmes first shift foundational decoding skills for Grade 1 learners, before relatively greater improvements on higher order domains of reading proficiency follow in Grade 2. In the first year of the EGRS coaching programme implementation (on Grade 1 learners), it only had a significant effect on one foundational domain of reading proficiency: phonological awareness (Cilliers et al., 2019: 35). In the second year when the evaluation cohort were in Grade 2, the estimated effect on foundational decoding skills like letter sounds and phonological awareness were smaller. In contrast, the impact on higher order decoding and reading skills (word recognition, non-word reading and paragraph reading) was significantly larger in the second year relative to the first. The dynamic impacts of the EGRS study thus align with the Funda Wande findings in suggesting that the acquisition of decoding and then higher order reading fluency and comprehension skills are sequential in nature.

The EGRS 1 coaching treatment impacts were clearer for boys than for girls at midline, with the treatment estimate equal to 0.19 s.d. for boys (significant at the p=0.05 level). In contrast, for the Funda Wande Grade 1 sample, if anything, the impact was greater for Grade 1 girls. The midline results from the EGRS 1 coaching intervention suggests that it helped the generally lagging boys catch up with the girls in their classrooms for reading comprehension. In the Funda Wande sample, the programme does eventually seem to help boys in intervention schools catch up to their girl counterparts, but only in Grade 2, and only after girls might have pulled further away during Grade 1

Both programmes performed similarly in the sense that no learners experienced negative treatment effects based on their baseline levels of reading proficiency. Programme impacts also did not vary with learners' relative rank for reading proficiency within their classrooms. However, the Funda Wande intervention seems more likely to have a positive benefit to learners in intervention schools across the distribution of baseline reading proficiency levels. Importantly, the Funda Wande programme had a clear positive effect on the weakest learners based on baseline reading proficiency, whereas this was not the

48. A comparison for the letter sounds sub-task in the two studies provides insights into the effects of the two programmes. Grade 1 leaners in both the EGRS I and Funda Wande control schools could only recognise about five or six letters correctly per minute at the start of Grade 1 (Taylor et al., 2016: 30). Control school learners' average letter recognition ability also grew at almost the same rate in status quo Grade1 learning environments in both studies, by 17 and 18 correct letter sounds per minute after one year in EGRS 1 and Funda Wande respectively. The point estimates of the programmes are 0.13 s.d. (p-value = 0.084) and 0.27 (p-value=0=01) respectively (Taylor et al., 2016: 42). As with the composite reading proficiency impact estimates, the confidence intervals of the respective programme impacts on letter sound recognition overlap. The more important insight is that the effect sizes translates to between two and six additional letter sounds per minute for intervention school Grade 1's, or between 12 percent to a third of a year of learning in comparison schools.

case in the EGRS 1 evaluation after either one or two years of programme implementation. These results suggest that a) it is students' absolute levels of reading proficiency that matters for this class of programme effectiveness, and b) there is some characteristic(s) of the Funda Wande programme that make effective for leaners across the distribution of initial reading proficiency levels.

At this stage, we can only speculate whether the differential impact on initially lower performing learners is attributable to differences in programme design. For example, the Funda Wande training materials places a specific focus on ongoing formative assessment and the urgency of remedial interventions targeted at students who fall behind (Funda Wande, 2018: 42-44). This is accompanied by the provision of baseline assessment booklets, for which the expectation is that they will be used to assist teachers in placing learners into ability groups for Group Guided Reading activities. The Funda Wande programme is also distinct in its emphasis on the affective response that learners have to reading. Amongst other things, it emphasises the importance of teachers understanding and seeking out the sources of what motivates their learners to read, developing strategies for establishing a culture of reading for enjoyment, and providing learners with reading challenges tailored to their level of proficiency (to facilitate sustained engagement and interest on the part of the learners). We are not able to test for these mechanisms at this stage. However, in-depth classroom observations should shed further light on the mechanisms at play.

With respect to the potential mechanisms at play in the EGRS I study, a combination of teacher surveys, document inspections and in-depth classroom observations indicate that teachers in the coaching arm felt more supported, their classrooms had more access to print resources, and treated teachers adhered more closely to the curriculum prescribed pacing, sequencing and coverage of reading materials and practices. Despite the coached teachers having a better knowledge of, and adherence to, curriculum prescribed routines, Cilliers et al. (2019) make a careful argument that the main driver of programme impacts is coached teachers being more likely to implement technically demanding instructional techniques (especially Group Guided Reading). In turn, their students are more likely to read materials at their proficiency levels, receive individualised attention when reading, and also spend more time actually using the reading materials provided (Cilliers et al., 2019: 22-25). This aligns with the findings on reported classroom practices of intervention schools in the Funda Wande intervention: the increased use of Vula Bula Graded readers, a shift away from communalised learning towards more individualised modes of reading instruction, and the increased propensity to make use of group guided reading specifically.

Another way to get a sense of the relative size of the programme impacts is to compare the results against benchmarks from "studies of studies": meta-analyses and systematic reviews. Conn (2017) provides a particularly relevant benchmark. Based on a meta-analysis of 66 experiments from across Sub-Saharan Africa, pedagogical interventions aimed at shifting teacher pedagogy and/ or instructional techniques yields a conservative estimated effect size of 0.228 s.d.. Kraft et al.'s (2018) perform a meta-analysis for 21 causal studies from North America focussed on teacher coaching type professional development programmes and estimate the impact for early Grade reading outcomes specifically. The pooled effect size estimate is 0.19 s.d. From other systematic reviews, McEwan's (2015) estimated effect size for teacher training programmes more broadly is 0.12 s.d. and Snilstveit et al. (2016) find a pooled estimate on structured pedagogy programmes of 0.23 s.d.

However, the available evidence suggests that there is more variation in effectiveness across teacher professional development programmes

than between different types/classes of educational interventions more broadly (Evans and Popova, 2016; McEwan, 2015). This is a recurrent theme across educational interventions: there is a lot of variation in the effectiveness within programme types and therefore a range of considerations that must be taken into account when interpreting the effect sizes of various education interventions[49](see Kraft, 2019). This is consistent with arguments to shift focus from questions of "what works", to "how it works"; i.e. that the most useful experimental studies are those trying to identify causal mechanisms behind a programme's impacts, or lack thereof (Deaton, 2010, Ludwig et al, 2011).

It is important to keep in mind at this stage that Funda Wande schools were screened before they were invited to participate in the programme, based on certain pre-defined selection eligibility criteria (refer to section 5.3). More precisely, all the schools in the sample are drawn from no-fee, quantile three public schools that have an isiXhosa LOLT and are located on the in- and around three urban school districts in the Eastern Cape. Other eligibility criteria include that schools should not suffer from severe overcrowding, nor have less than 20 learners per Grade. These sample criteria guided the process of selecting which characteristics by which to test for differential impacts, and it should also be kept in mind when considering the extent to which the sample's heterogeneity on certain characteristics are representative of the broader population of schools in South Africa.

The external validity consideration relevant here is whether those schools who agree to participate and/or were selected to participate by the programme's implementers are different from the rest of the population (Banerjee et al, 2017:80; Kraft, 2019:12). "Site-selection bias" refers to the phenomenon where sub-groups and/or locations for the intervention are chosen such that marginal returns are expected to be particularly large (Allcott, 2015; Banerjee et al, 2017:81). Both the EGRS programmes, as well as the Funda Wande programme, are implemented in low-resource, no fee schools in provinces with academic outcomes generally below the national average (Kotze et al., 2019: 206; Taylor et al., 2017: 16). These contexts arguably provide more fertile grounds for greater learning improvements if the programme class' effectiveness diminishes with the pre-existing teacher capacity and/or student learning levels within schools

On the other hand, there is a lot of homogeneity in poor learning outcomes for learners in schools across the bottom four wealth quintiles of South African schools (Spaull an Kotze, 2015; Spaull, 2013), as well as strong correlation between school wealth, -location, the socioeconomic status of learners and the language of instruction (Spaull, 2013). It seems plausible that the underlying mechanisms that make these structured pedagogy programmes effective could be similarly applicable across low-resource, low capacity South African schools more generally[50]. If effect sizes therefore hold with some consistency across provinces, for schools with different languages of instruction, both rural/urban locations, etc., then the effects estimated in the low resource environments where the programmes are evaluated are indeed the parameter of interest to policy-makers.

The context dependence of these structured pedagogy programmes is best assessed by replication of the same/similar interventions (i.e. functioning on the same underlying rationale) across different contexts (Banerjee et al, 2017:96). In the South African policy space this could imply replication across different provinces, time periods, partnering provincial bureaucracies, implementing institutions and languages of instruction. To date, the state of evidence from the EGRS I (Cilliers et al., 2019a, 2019b), EGRS II (Kotze et al., 2019) and Funda Wande coaching interventions suggest that structured pedagogy programmes can be effective in more than one province (the North West, Mpumalanga and the Eastern

49. For example, something that is often neglected in the comparison of the relative magnitude of programme impacts is the important role played by the choice of outcome measure used in determining programme effectives (i.e. which outcomes are measured, when they are measured and how they are measured - Kraft, 2019: 11-12). For example, teacher coaching interventions have much larger effects on more proximal outcomes (like shifting teachers' instructional practice - 0.47 SD) than on downstream students' achievement (0.18 SD), whilst literacy outcomes are generally easier to shift than achievements in maths (Kraft, Blazar, & Hogan, 2018).

50. For example, school wealth is a very strong predictor a learner's probability of learning to read by Grade 4 (Spaull and Pretorius, 2019:156). Based on PIRLS 2016 data, Spaull and Pretorius (2019:156) estimate that "(t)he average child in the poorest 75% of schools has a five times higher probability of not learning to read than of learning to read (85% compared to 15% respectively)." Whilst there are a few outlier children in in no-fee (quintile 1-3) schools who manage to learn to read, there are fewer, if any outlier schools (where the majority of learners learn to read in spite of the school's limited resources).

It seems plausible that the underlying mechanisms that make these structured pedagogy programmes effective could be similarly applicable across low-resource, low capacity South African schools more generally

Cape) and in more than one language of instruction (Setswana Home language, English as first additional language and isiXhosa Home language). This lends further external validity to earlier experimental evidence that on-site coaching interventions can shift instructional practice and improve early learning outcomes in low-resource South African schools.

Nevertheless, the variation in effectiveness of structured pedagogic interventions observed in the literature are not only a function of the programme type's applicability and transferability to different contexts (referring to the programmes external validity- see Bates and Glennerster, 2017; Pritchett and Sandefur, 2015). One specific programme is also only one realisation in a class of possible iterations under the same umbrella class of intervention ("teacher incentives", "providing textbooks", "coaching", etc.). The latter, which Nadel and Pritchett (2016) call a class of programme's design space, refers to the multiple design elements and potentially many possible choices within each of those design elements. Together, the role of contextual factors and the specific design- and implementation details of programmes could explain why multiple school based interventions that have succeeded at the pilot and proof of concept stage (often implemented with high fidelity by committed non-governmental organisation (NGO) implementers and at a smaller, more manageable scale) often fail to reproduce positive affects when replicated at scale (as is generally required by governments) (see Bold et al., 2018); or why the same program implemented by the same actor across different countries yield different outcomes (Lucas et al., 2014). Subsequent evaluation reports on the Funda Wande intervention should therefore include detailed programme implementation information, both in terms of what the goals for various programmatic components were and how well the goals were achieved in practice (i.e. implementation fidelity).

Detailed knowledge of programme components, how they fit together in a theory (i.e. a theory of change), and how the intervention implementation plays out in practice (implementation fidelity) are important components of understanding how and why a programme might be effective in different contexts. The Funda Wande programme has extensive documentation available of the range of programmatic inputs, why these inputs are required, and how they are interlinked parts of the bundled intervention (Funda Wande, 2018). An unanswered question at this stage is how the respective inputs are perceived and used in practice, and to what extent the respective programme inputs are indeed complementary.

It is clear from the existing evidence on teacher focussed interventions that furnishing teachers with either knowledge or resources in isolation is not enough (Glewwe et al., 2009, 2014; Kennedy, 2016). Instead, a holistic bouquet of complementary materials and support is provided to teachers and

Given the success of the Funda Wande programme at this early stage, another consideration for the interpreting the generalisability of programme impacts is to consider how they might look at scale.

the learners in their classrooms. For example, Piper et al. (2018) examined the relative impact on learning, as well as the cost-effectiveness, of the constituent components of a similar programme (Tusome) in Kenya. The intervention differentially provided a combination of initial teacher training, follow-up coaching, literacy materials in a 1:1 ratio per student, and structured teachers' guides. There results show that the entire package combined resulted in much greater impacts than a) only books and training or b) teacher training alone

Similarly, Kerwin and Thornton (2019) evaluate a lower-cost version of a similar mother-tongue early literacy programme that proved highly effective in Uganda, but when implemented at a relatively small scale and by a high performing non-governmental organisation (NGO). The original programme provided a version of the general structured pedagogical intervention bouquet: detailed lesson plans, teacher training with subsequent monitoring and support, mother tongue readers and other classroom materials. The reduced cost version of the model was designed to emulate programme modifications often undergone to make an effective pilot programme more amenable to implementation at scale and by government implementers. Besides cutting seemingly inessential but expensive classroom materials (the provision of classroom clocks and writing slates), the major cost-saving changes were to the method of teacher training (moving to a cascade, "train-the-trainer" model) and reducing the number of follow-up support visits to teachers (now done by trained government officials and not NGO implementers). These changes cut programme costs by 60%. However, it also reduced the impacts of the programme across a range of learning outcomes to such an extent that the reduced cost version of the programme was 40% less cost-effective than the pilot programme originally implemented.

From the perspective of the Funda Wande evalutation, this implies that future rounds of programme expansion and -evaluation could greatly improve the state of knowledge on the importance of individual programme components. This would entail an assessment of the extent to which inputs are complementary, by assessing the whether individual Funda Wande programme components or combinations of sub-sets of them contribute to shifting teacher practice and learner outcomes.

Given the success of the Funda Wande programme at this early stage, another consideration for the interpreting the generalisability of programme impacts is to consider how they might look at scale. From an evaluation perspective, this has a few implications. One is investigating whether the programme has any spillover effects. For example, a programme might lead to a different set of behaviours (and subsequent effects on learning outcomes) when only some schools receive the treatment, compared to nation-wide implementation where all schools

receive the programme. A relevant example to the structured pedagogy class of interventions is the potential usefulness of informational spillovers and shared learning from teachers forming learning communities (often amongst small clusters of near-by schools) where best practices are shared (Fleisch et al., 2017: 10). At scale, these spillovers could well lead to even greater instructional changes and learning gains than in evaluated programmes where treated schools are further apart due to cluster/school level randomisation. From EGRS I, Cilliers et al. (2019b) do find positive spillovers on untreated teacher's pedagogical knowledge among teachers not exposed to the intervention within the treated schools[51], but none between schools (Taylor et al., 2017: 117-119).Another consideration for taking a programme to scale is its cost-effectiveness. This requires detailed data on both the costs of various programmatic inputs and subsequently calculating returns to the various domains of reading proficiency per rand spent. Understanding how cost effective the programme (or some version thereof) might be at scale provides another motivation for considering the role of individual- or sub-sets of programme inputs separately in future. For example, if only providing HOD training (through the -year part-time accredited course) and the Funda Wande LTSM resources are sufficient to shift learning outcomes and do so in a more cost effective manner, this would arguably provide a model of programme implementation that is more amenable to operating at scale within a national level public education system. At the very least, however, a detailed costing exercise of the Funda Wande programme will be conducted on implementation of the programme as is for the next round of evaluation.

Current evidence from the EGRS 1 study provides a useful framework for thinking about cost-effectiveness considerations in the South African context. It assesses the effectiveness of two different teacher professional development delivery modalities (centralised training versus in-person coaching) and how these relate to the persistence of programme effects over the longer run. Even though the EGRS I coaching intervention is more expensive in absolute terms (R557 / 42.91 USD versus R397 / 30.58 US per pupil[52]), it is more cost effective in terms of learning gains per USD spent when evaluated over two years. There is an estimated 0.57 s.d. increase in reading proficiency annually and per 100 USD spent per pupil in the coaching arm, compared to 0.39 s.d. in the training arm. When evaluated over a longer time-span, however, it is unclear whether the coaching programme is still more cost-effective (Cilliers et al., 2019b:17).

Nevertheless, the coaching intervention did perform relatively better on certain key impact measures of the longevity of programme impacts. Subsequent follow-ups of the EGRS I study found that the initial shifts in teachers' pedagogical knowledge, resource use and subsequent student learning improvements generally persist up to two years after the intervention for both trained and coached cohorts. However, only teachers in the coaching arm continued making use of the programme acquired instructional techniques. The effect of the programme on subsequent cohorts of learners was also more persistent in the coaching arm, with coached teachers being the only treatment arm to have significant positive impacts on subsequent cohorts of students' learning one year after receiving the intervention (Cilliers et al., 2019b). Cilliers et al. (2019b) suggests that sustained and meaningful change in teacher practices, which are necessary for improved better student learning outcomes, requires some form of ongoing in-classroom support, monitoring and feedback for teachers, like that offered by the coaches in the intervention

Viewed in conjunction with preliminary evidence from a subsequent first additional language (English) EGRS II programme, which compares outcomes on solely face-to-face versus e-coaching based teacher training (Kotze et al, 2019), it is not

51. The authors argue that within-school spill-over effects are due to social learning and/or the use of educational materials provided (Cilliers, 2019b:4).

52. This is based on variable costs only (see discussion in Taylor et al.,2019: 92).

yet clear which method of teacher professional development is the most cost effective. Also noteworthy is that implementing programmes at national scale might change the cost effectiveness of the respective programmes, following from changes to prices and capacity constraints as programme inputs are required in much larger quantities (i.e. general equilibrium effects). For example, a shortage of well-qualified and – trained coaches could potentially both decrease the impacts of a teacher coaching programme at scale, and simultaneously increase its initial implementation costs (through the initial investment required to train the larger number of coaches required to implement the programme nationally[53]). For the Funda Wande programme specifically, an important consideration at scale will be whether components like online teacher resources and the flash disc of multi-media materials can improve the cost effectiveness of the programme.

A final consideration for thinking of how a programme might look at scale is the political reactions that it will encounter, be they supportive or in opposition. In the South African context, there is arguably a lot of support for structured pedagogic interventions – driven to a large extent by the successful interventions by the DBE itself. Positive evaluation results and teacher feedback also feed into the continued support for the interventions from various stakeholders at this stage. However, political reactions might well change with the scale at which the programme is implemented and thus the extent to which it impacts on the interests of various role-players (Banerjee et al, 2017: 78). This is not something that programme implementers have much direct control over, but it is worth being cognisant of in the early stages of implementation as it can result in implementation challenges of a different kind. Thoroughly assessing teacher sentiment toward the programme and it's respective components might provide a useful indicator at the early stages of implementation.

In sum, there are a few important considerations for programmes like Funda Wande to operate at scale in future. The first lie in further understanding the persistence of programme impacts and how this relates to cost effectiveness (Cillers et al, 2019b). Another challenge is settling on implementation modalities that are amenable to operating at scale in a cost-effective manner. This includes the programme remaining politically feasible and being effective when implemented within a national-level, public-school education system (most likely by government bureaucrats, and not NGO staff). Given the extent to which the cost-effectiveness of programmes can differ with context- and implementation scale changes (Tulloch, 2019), the Funda Wande programme will provide another important data-point on the cost-effectiveness of mother-tongue, structured pedagogical interventions in the South African context. It also provides another puzzle-piece in creating a fuller picture on the class of programmes' effectiveness across different contexts.

53. See discussion in Banerjee et al. (2017:75-77) on general equilibrium effects and how they relate to scaling (and evaluating) successful pilot programmes.

9. Next Steps

9. Next Steps

Expansion of scope of intervention and evaluation to include numeracy:

During the course of 2019, Funda Wande met with the 30 school principals and School Management Treams involved in the intervention as well as the ECDoE district officials in Sarah Baartman, Nelson Mandela Bay and Buffalo City. One of the points that came out of those interactions was that teachers, HODs and district officials all requested an extension of the Funda Wande programme to provide support for Grade R-3 mathematics.

In response to this and with generous support from donors, the scope of the programme has been expanded to include a mathematics component. Bala Wande: Calculating with Confidence is headed up by Ingrid Sapire (Wits) who was also appointed by the Minister of Basic Education Ms Angie Motshekga to be the Chairperson of the Ministerial Task Team to develop the Framework for Teaching Mathematics for understanding (TMU).

Given that all Foundation Phase teachers teach both literacy and mathematics, the coaches will now provide support for both of these subjects, i.e. it is the same schools with the same teachers and learners. The Bala Wande intervention will begin with Grade 1 in 2020 and expand into Grade 2 and 3 in 2021 and 2022 respectively.

Planned future activities for the evaluation:

The main research activities planned for the next three years (2020 – 2022) are as follows:

1. September 2020: Conduct classroom observations and semi-structured teacher interviews in a small-subsample of schools. This qualitative data will generate valuable insights into the take-up of Funda Wande techniques in the classroom. Such observations will focus on precisely how teachers implement the programme, and why teachers take up the intervention in particular ways.

2. October/November 2020: Conduct the third round of data collection (midline II) in all 59 schools. The evaluation team will assess the same learners (Cohort A and B) that were assessed at baseline and midline I. The team will also assess randomly selected Grade one learners (Cohort C). Assessments will focus on literacy and numeracy skills.

3. October/November 2021: Conduct the fourth round of data collection (midline III) in all 59 schools. The evaluation team will assess learners in Cohort A and C.

4. October/November 2022: Conduct the final round of data collection (endline) in all 59 schools. The evaluation team will assess learners in Cohort C.

10. Conclusion

10. Conclusion

The midline results for the Funda Wande in-service teacher coaching intervention indicate that the programme is effective in improving Grade 1 and 2 learners isiXhosa home language reading outcomes after one school year of programme exposure. Partnering with the Eastern Cape Department of Education (ECDoE), the intervention bouquet of providing home language text resources for learners and classrooms, curriculum aligned lesson plans, assessment booklets and online pedagogic resources for teachers, accompanied by in-classroom coaching support, feedback and HOD training led to a 0.17 s.d. impact on learners' reading proficiency.

In terms of the amount of learning that took place in comparison school 'business as usual' learning environments, this effect translates to between 20 to 27 percent of a year's worth of learning for Grade 2 learners, and 33 to 58 percent of a year's learning for Grade 1 learners. Dependent on the outcome measure used, the programme impacts therefore range roughly between one and two terms of learning in status quo schooling environments in these three Eastern Cape districts.

The programme effects are positive across all sub-tasks for all the sub-domains of reading proficiency that could be reliably measured. For Grade 1 learners, intervention impacts were the largest on foundational decoding skills - correctly identifying letter sounds and being able to manipulate phonemes. At this early stage of Grade 1 leaners' development trajectories, these are the skills that are required to decode words, read more fluently and eventually read for meaning. Subsequently, the impacts on further downstream higher order reading comprehension skills are only detectable for Grade 2 learners. Viewed together with other results from the recent literature, these results support the idea that learners require a range of foundational literacy abilities before they can read with some level of speed and accuracy (i.e. fluency), and in turn, they need to read with a certain minimum level of fluency in order to comprehend what they are reading.

A particularly encouraging finding from a policy perspective is that the intervention has fairly consistent positive impacts for learners across the distribution of baseline reading proficiency. Programme impacts also did not vary with of learners' relative rank for reading proficiency within their classrooms. Previous research suggests that improving reading outcomes for learners with the lowest levels of foundational reading skills in an absolute sense is particularly challenging (Cilliers et al., 2019). A potentially related

finding is the suggestive evidence that the programme helps boys in treatment schools catch up with their girl counterparts, but only in Grade 2 and with the extent of catch-up being contingent on the boys' baseline levels of reading proficiency.

At this stage only suggestive results are presented for the potential mechanisms at play. Evidence across more than one indicator suggests that teachers in intervention schools are more likely to *a)* be more attuned to the actual reading proficiency levels of the learners in their class (both in terms of whether learners are at the top or the bottom of the distribution and how the class performs overall), *b)* to make use of material resources provided more often, and *(c)* to use instructional techniques that have previously shown to facilitate more individualised forms of actual learner reading practice and -teacher feedback (Cilliers et al., 2019). Future rounds of assessments and in-depth qualitative classroom observations will delve deeper into both the potential mechanisms at play, as well as the potential characteristics of the Funda Wande intervention that result in it being effective in shifting learning outcomes for leaners across the distribution of reading proficiency levels (and for learners with the lowest levels of reading proficiency in particular).

Other unanswered question at this stage relate to the details that would allow one to compare the absolute- and cost-effectiveness of the programme to similar interventions in the literature. Primarily, these considerations relate to in-depth programme implementation details (like the hours of exposure to the programme per teachers) and carefully costing of the intervention over its first two years of implementation.

The results here add to the growing body of evidence that makes a strong case for the crucial complementary role of high-quality teacher coaching and continuous follow-up support in programmes that focus on shifting teachers' instructional practice. Similar to other programmes in Kenya, Uganda and South Africa (Piper et al., 2014, 2018; Kerwin and Thornton; 2019, Cilliers et al., 2019a), the Funda Wande intervention has shifted learning outcomes through combining material provision, a structured sequence of lessons, alignment around some central curriculum, and supporting teachers in "learning by doing" through teacher professional development support. For the package of interventions to be successful, indications are that some degree of coaching, monitoring and feedback focussed on specific instructional techniques, lesson planning, the effective use of newly provided materials; and the implementation of more technically demanding teaching techniques (such as group-guided reading) are all contributing constitutive components.

Given the consistent positive impacts found for the structured pedagogy programmes that have been assessed in the South African context (across different provinces, time periods, partnering provincial bureaucracies, implementing institutions and languages of instruction.) - these programmes are arguably past the proof of concept stage and have begun to establish a set of generalizable lessons. However, many questions that remain are related to implementing a version of these programmes at scale: 1) what role do the individual inputs and combinations of them play in driving programme impacts (for example, the provision of home language resources and instruction), 2) how cost effective are different iterations of the class of intervention and 3) how would the relation of programme costs and benefits change if it were implemented at scale within a national level public education system in future.

11. References

Allcott, H. 2015. Site Selection Bias in Program Evaluation. Quarterly Journal of Economics, 130(3): 1117–65.

Bates, M. A., and Glennerster, R. 2017. The generalizability puzzle. Stanford Social Innovation Review, Summer, 50–54.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shortland, M and Walton, M. 2017. From proof of concept to scalable policies: Challenges and solutions, with an application. Journal of Economic Perspectives, 31(4): 73-102.
Bau, N. and Das, J. 2019. Teacher Value-Added in a Low-Income Country. American Economic Review: Economic Policy, forthcoming.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A. and Sandefur, J. 2018. Experimental evidence on scaling up education reforms in Kenya. Journal of Public Economics, 168: 1-20.

Bold, T., Filmer, D., Martin, G., Molina, E., Rockmore, C., Svensson, J. and Wane, W. 2017. What do teachers know and do? does it matter? evidence from primary schools in Africa. World Bank Policy Research Working Paper 7956. Washington, D.C.: World Bank.

Brunette, T., Piper, B., Jordan, R., King, S. and Nabacwa, R. 2019. The Impact of Mother Tongue Reading Instruction in Twelve Ugandan Languages and the Role of Language Complexity, Socioecoomic Factor, and Program Implementation. Comparative Education Review, 63(4): 591-612.

Bruns, B. and Luque, J., 2014. Great teachers: How to raise student learning in Latin America and the Caribbean. Washington, D.C.: World Bank.

Chetty, R., Friedman, J. and Rockoff, J., 2014. Measuring the impacts of teachers II: Teacher valueadded and student outcomes in adulthood. American Economic Review, 104(9): 2633-79.

Cilliers, J., Fleisch, B., Prinsloo, C., and Taylor, S. 2019a. How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. Journal of Human Resources, in press.

Cilliers, J., Fleisch, B., Kotze, J., Mohohlwane, M. and Taylor, S. 2019b. The sustainability of early Grade education interventions: Do learning gains and improved teacher practices persist? Gui2de working paper series no. 6. Georgetown University Initiative on Innovation, Development and Evaluation (Gui2de). Available at: https://repository.library.georgetown.edu/handle/10822/1055275. [Date of access: 15 October 2019].

Conn, K. M. 2017. Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. Review of Educational Research, 87(5), 863–898.

Das, J., Dercon, S., Habyarimana, J, Krishnan, P., Muralidharan, K. and Sundaraman, V. 2013. School Inputs, Household Substitution, and Test Scores. American Economic Journal: Applied Economics 2013, 5(2): 29–57.
Deaton, A. 2010. Instruments, Randomization, and Learning about Development. Journal of Economic Literature, 48(2), 424–45.

Evans, D. K., & Popova, A. 2016. What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. The World Bank Research Observer, 31(2), 242–270.

Fleisch, B., Schöer, V., Roberts, G., and 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. and 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.

Glewwe, P; Hanushek, E. A.; Humpage, S; Ravina, R. 2014. School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010. In Glewwe, P. (ed.). Education policy in developing countries. Chicago, IL: University of Chicago Press. pp. 13-64.

Glewwe, P., Kremer, M., and Moulin, S. 2009. Many Children Left Behind? Textbooks and Test Scores in Kenya. American Economic Journal: Applied Economics, 1(1): 112-35.

Hanushek, E. and Rivkin, S. 2010. Generalizations about using value-added measures of teacher quality. American Economic Review, 100(2): 267-71.

Howie S, Combrink C, Roux K, Tshele M, Mokoena G, Palane N. 2018. Progress in international reading literacy study 2016: South African children's reading literacy achievement. Pretoria, South Africa: Centre for Evaluation and Assessment, University of Pretoria.

Hoadley, U. 2018. Pedagogy in poverty: Lessons from twenty years of curriculum reform in South Africa. Abingdon/New York: Routledge.
Hoadley, U. 2012. What do we know about teaching and learning in South African primary schools? Education as Change, 16 (2):187–202.

Kennedy, M. 2016. How does professional development improve teaching? Review of Educational Research, 86(4), 945–980.

Kerwin, J.T. and Thornton, R.L. 2019. Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures.

Kerwin, Jason and Thornton, Rebecca L., Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures (January 30, 2018). Available at: https://ssrn.com/abstract=3002723. [Date of access: 15 October 2019].

Kotze, J., Fleisch, B. & Taylor, S. 2019. Alternative forms of early Grade instructional coaching: Emerging evidence from field experiments in South Africa. International Journal of Educational Development, 66(1), 203-213.

Kraft, M. A., Blazar, D., & Hogan, D. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. Review of Educational Research, 88(4), 547-588.

Kraft, M.A. 2019. Interpreting Effect Sizes of Education Interventions. Annenberg Institute EdWorkingPaper: 19-10. Brown University. Available at: http://www.edworkingpapers.com/ai19-10 . [Date of access: 15 October 2019].

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. 2014. Improving early-Grade literacy in East Africa: Experimental evidence from Kenya and Uganda. Journal of Policy Analysis and Management, 33(4), 950–97

Ludwig, J., Kling, J. R., & Mullainathan, S. 2011. Mechanism Experiments and Policy Evaluations. The Journal of Economic Perspectives, 25(3), 17–38.

McEwan, P. J. 2015. Improving Learning in Primary Schools of Developing Countries: A MetaAnalysis of Randomized Experiments. Review of Educational Research, 85(3), 353–394.

Menendez, A., and Ardington, C. 2018. Impact Evaluation of USAID/ South Africa Story Powered School Program – Baseline. Technical Report. Available at: https://nalibali.

org/research. [Date of Access: 5 October 2019].

Motala, S, and Carel, D. 2019. Education Funding and Equity in South African Schools. In Spaull, N. and Jansen, J.D. (eds.) South African Schooling: The Enigma of Inequality. Springer. Chapter 4.

Mullis I.V.S, Martin, M.O., Foy, P.., Hooper, M. 2017. PIRLS 2016: International Results in Reading. International Association for the Evaluation of Educational Achievement.

Nadel, S., and Pritchett, L. 2016. Searching for the Devil in the Details: Learning About Development Program Design. Working Paper No. 434. Center for Global Development. Available at: https://www.cgdev.org/publication/searching-devil-details-learning-about-development-program-design-working-paper-434 . [Date of Access: 1 December 2019].

NEEDU. 2013. NEEDU National Report 2012: The State of Literacy Teaching and Learning in the Foundation Phase. National Education and Evaluation Development Unit. Technical report. Pretoria: Department of Basic Education.
Ouellette, G.P. and Haley, A. 2013. One complicated extended family: The influence of alphabetic knowledge and vocabulary on phonemic awareness. Journal of Research in Reading, 36(1): 29–41.

Piper, B. and Korda, M. 2011. Early Grade reading assessment (EGRA) plus: Liberia: Program evaluation report. Research Triangle Park, NC: RTI International.

Piper, B., Kwayumba, D., Oyanga, A., & Jepkemei, E. 2015. The Primary Math and Reading (PRIMR) Initiative endline impact evaluation on the DFID Kenya Rural Expansion Programme. Prepared for DFID Kenya under contract 202657-108. RTI International, Research Triangle Park, NC

Piper, B., Zuilkowski, S.S., Dubeck, M., Jepkemei, E., & King, S. J. 2018. Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. World Development, 106, 324–336.

Piper, B., Zuilkowski, S. S., and Mugenda, A. 2014. Improving reading outcomes in Kenya: First-year effects of the PRIMR initiative. International Journal of Educational Development, 37: 11–21.

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 (USAID).

Pretorius, E.J., and Spaull, N. 2016. Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. Reading and Writing, 29:1449–1471.

Pritchett, L., and Sandefur, J. 2015. Learning from Experiments when Context Matters. American Economic Review 105(5): 471–75.

Sailors, M., Hoffman, J. V., Pearson, P.D., Beretvas, S.N., and Matthee, B. 2010. The Effects of First- and Second-Language Instruction in Rural South African Schools. Bilingual Research Journal, 33 (1): 21–41.

Snilstveit, B., Stevenson, J., Menon, R., Philips, D., Gallagher, E., Geleen, M. 2016. The impact of education programmes on learning and school participation in low- and middle-income countries. Systematic review summary 7. London: International Initiative for Impact Evaluation (3ie).

Snow, C. 2017. Early literacy development and instruction: An overview. In The Routledge

international handbook of early literacy education: A contemporary guide to literacy teaching and interventions in a global context, eds. Natalia Kucirkova, Catherine E. Snow, Vibeke Grøver, and Catherine McBride-Chang, 5-13. Abingdon, Oxon; New York, NY: Routledge.

Spaull, N. and Kotze, J. 2015. Starting Behind and Staying Behind in South Africa. The Case of Insurmountable Learning Deficits in Mathematics. International Journal of Educational Development, 41:13–24.

Spaull, N., Pretorius, E. and Mohohlwane, N. 2020. Investigating the comprehension iceberg: Developing empirical benchmarks for early-Grade reading in agglutinating African languages. South African Journal of Childhood Education, 10(1): 1-14.

Spaull, N. and Pretorius. 2019. Falling at the first hurdle: Early Grade reading in South Africa. In Spaull, N. and Jansen, J.D. (eds.) South African Schooling: The Enigma of Inequality. Springer. Chapter 8. pp 147-268.

Spaull, N. 2019. Equity: A Price Too High to Pay. In Spaull, N. and Jansen, J.D. (eds.) South African Schooling: The Enigma of Inequality. Springer. Chapter 1. pp 1-24.

Taylor, S., Cilliers, J., Prinsloo, C., Fleisch, B. and Reddy, V. 2019. Improving Grade Reading in South Africa. New Delhi: International Initiative for Impact Evaluation (3ie). [Grantee Final Report].

Taylor, S., Cilliers, J., Prinsloo, C., Fleisch, B. and Reddy, V. 2017. The Early Grade Reading Study: Impact Evaluation after Two Years of Interventions. Department of Basic Education: Pretoria. [Technical Report].

Taylor, S., and 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.

Tulloch, C. 2019. Taking intervention costs seriously: a new, old toolbox for inference about costs. Journal of Development Effectiveness, 11(33):273-287.

Van der Berg S, Spaull N, Wills G, Gustafsson M, Kotze J. 2016. Identifying binding constraints in education. Report commissioned by the South African Presidency and funded by the European Union's Programme to Support Pro-Poor Policy Development (PSPPD) initative. Pretoria.

World Bank. 2018a. Learning to Realize Education's Promise. World Development Report 2018, Washington DC.

World Bank. 2018b. An Incomplete Transition: Overcoming the Legacy of Exclusion in South Africa. Systematic Country Diagnostic, Washington DC.

What Works Clearinghouse. 2020. What Works Clearinghouse Standards Handbook, Version 4.1. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/handbooks

Zuilkowski, S. S., & Piper, B. 2017. Instructional coaching in Kenya: Supporting teachers to improve literacy outcomes. In M. Akiba & G. K. LeTendre (eds.), International handbook of teacher quality and policy. New York, NY: Routledge. pp. 505–516.

12. Appendix

Table A1: Testing for differential attrition		
	Learner Attrite	Teacher Attrite
Treatment	0.030 (0.032)	0.035 (0.053)
Control Attrition	0.046	0.069
Observations	1,187	1,187
R-squared	0.054	0.162
Strata FE	YES	YES

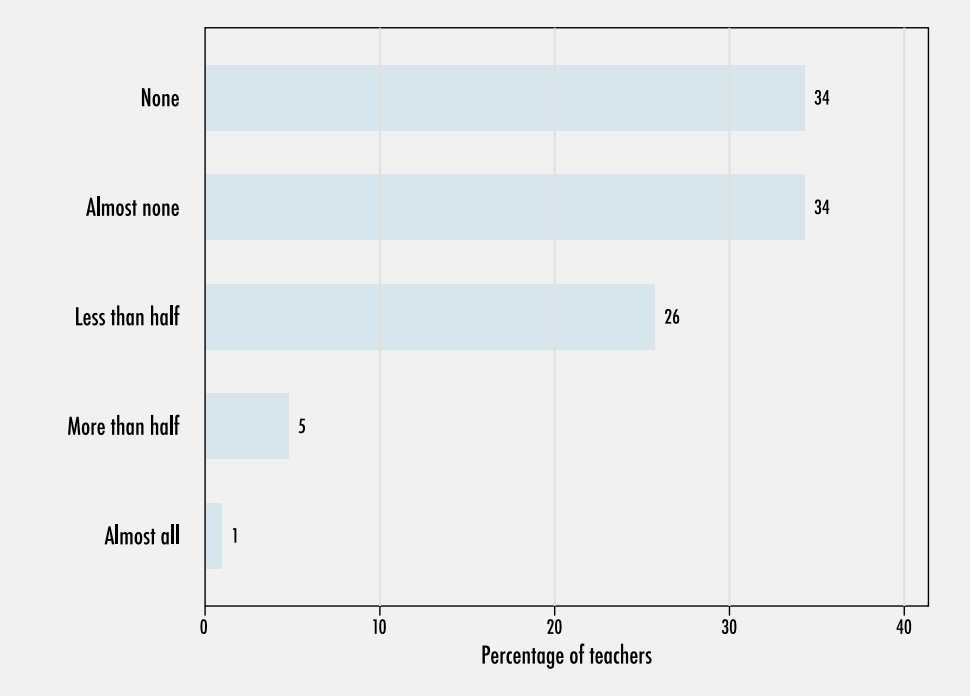
Table A2. Midline equivalence learner baseline test score, characteristics and home assets (both Grades)							
	Treatment	Control	p-value			Pooled	Effect
	Mean	s.e.	Mean	s.e.		s.d.	size
Common tasks							
Letter Sounds per minute	17,31	19,78	17,43	19,08	0,94	19,43	0,01
Digraphs and Trigraphs per minute	5,32	11,49	4,76	10,76	0,58	11,13	0,05
Productive Listening Comprehension	3,12	1,49	3,10	1,52	0,83	1,50	0,02
Receptive Listening Comprehension	9,63	0,74	9,50	0,98	0,13	0,87	0,15
Phonemic Awareness	2,87	2,51	2,97	2,52	0,65	2,52	0,04
Expressive Vocabulary	11,28	3,80	11,69	4,05	0,32	3,93	0,10
Write letters	2,70	1,93	2,89	1,80	0,26	1,87	0,10
Grade 1 only tasks							
Word Choice	1,73	1,74	1,90	1,59	0,37	1,67	0,10
Rapid Automatized Naming	36,99	12,00	37,60	11,62	0,61	11,80	0,05
Write your name	4,55	1,00	4,59	0,80	0,65	0,90	0,05
Copy a word	4,41	1,31	4,32	1,36	0,48	1,34	0,07
Grade 2 only tasks							
CVCV Words per minute	10,40	12,65	9,61	11,86	0,65	12,26	0,06
Familiar Words per minute	7,16	9,00	6,67	8,62	0,69	8,81	0,05
Oral Reading Fluency	7,74	9,92	7,47	9,14	0,84	9,53	0,03
Reading Comprehension	7,58	3,36	7,14	3,16	0,38	3,26	0,13
Vocabulary	3,14	2,39	3,23	2,24	0,78	2,32	0,04
Sentence Comprehension	4,54	4,47	4,63	4,34	0,88	4,40	0,02
Write words	14,03	7,00	14,63	6,46	0,57	6,73	0,09
Learner characteristics							
Grade 1	0,49	0,50	0,50	0,50	0,28	0,50	0,01
Grade 2	0,51	0,50	0,50	0,50	0,28	0,50	0,01
Female	0,51	0,50	0,49	0,50	0,33	0,50	0,05
Age in months	6,55	0,93	6,46	0,91	0,21	0,92	0,09
Height for age z-score	-0,37	1,04	-0,38	0,98	0,91	1,01	0,01
Household assets							
Books other than schoolbooks to read at home	0,33	0,47	0,38	0,49	0,09	0,48	0,12
Radio	0,71	0,46	0,72	0,45	0,73	0,45	0,02
Television	0,94	0,23	0,94	0,24	0,79	0,23	0,02
Computer	0,30	0,46	0,34	0,47	0,34	0,47	0,07
Toilet	0,59	0,49	0,62	0,49	0,54	0,49	0,05
Vehicle	0,50	0,50	0,53	0,50	0,57	0,50	0,04

Table A3: Treatment effects by subtask				
	Effect size (s.d.)	s.e.	p-values	
			Regression	Random. Inference
Both Grades				
Composite Score	0,17	0,05	0	0
Letters p/m	0,22	0,06	0,00	0
Di-/ Trigraphs p/m	0,16	0,06	0,01	0,02
CVCV Words p/m	0,14	0,05	0,01	0,04
Familiar Words p/m	0,17	0,05	0,00	0
Oral Reading Fluency	0,14	0,05	0,01	0,02
Reading Comp. I	0,11	0,06	0,09	0,13
Productive Listening	0,17	0,06	0,01	0
Vocabulary	-0,07	0,12	0,56	0,34
Phonemic Awareness	0,17	0,06	0,01	0,01
Grade 2				
Composite Score	0,16	0,04	0	0
Letters p/m	0,19	0,06	0	0
Di-/ Trigraphs p/m	0,17	0,06	0,01	0,01
CVCV Words p/m	0,15	0,05	0	0,01
Familiar Words p/m	0,19	0,05	0	0
Oral Reading Fluency	0,14	0,05	0,01	0,03
Reading Comp. I	0,13	0,06	0,03	0,09
Productive Listening	0,05	0,06	0,39	0,26
Phonemic Awareness	0,15	0,08	0,07	0,04
Oral Reading Fluency II	0,07	0,05	0,13	0,09
Reading Comp. II	0,12	0,06	0,03	0,02
Grade 2 Composite Score	0,15	0,04	0	0
Grade 1				
Composite Score	0,21	0,09	0,02	0,01
Letters p/m	0,27	0,09	0,01	0
Di-/ Trigraphs p/m	0,19	0,08	0,03	0,08
CVCV Words p/m	0,14	0,08	0,1	0,08
Familiar Words p/m	0,17	0,08	0,03	0,05
Oral Reading Fluency	0,16	0,08	0,04	0,06
Reading Comp. I	0,15	0,09	0,12	0,21
Phonemic Awareness	0,28	0,09	0	0,01
Productive Listening	0,2	0,08	0,02	0,04
Composite Score	0,21	0,09	0,02	0,01

Table A4: Ordered logit results for treatment effect on teacher ability to identify most/least proficient readers based on objective class rank		
	(1) Identified "Most proficient"	(2) Identified "Least proficient"
Treatment	-0.665**	-0.423
T=0 p-value	0.050	0.293
Observations	103	106

Table A5: Sub-task midline raw score distributions, by Grade and treatment status																			
	Control									Treatment									
Task	N	% zero	Mean	s.d.	Mean (excl. 0)	Percentiles				N	% zero	Mean	s.d.	Mean (excl. 0)	Percentiles			
						25th	50th	75th	max						25th	50th	75th	max
Grade 1																		
Letters p/m	279	8%	24,3	(18,5)	26,3	8	22	39	76	276	4%	28,6	(20,0)	29,7	10	27	43	84
Di-/ Trigraphs p/m	279	58%	6,2	(10,6)	14,7	0	0	9	55	276	50%	8,5	(13,7)	17,0	0	1	12	71
CVCV Words p/m	279	52%	6,3	(9,5)	13,2	0	0	10	40	276	54%	7,5	(11,3)	16,1	0	0	14	45
Familiar Words p/m	279	55%	4,2	(6,7)	9,4	0	0	6	27	276	57%	5,4	(8,2)	12,4	0	0	9	32
Oral Reading Fluency	279	56%	4,5	(7,6)	10,3	0	0	6	29	276	53%	5,8	(9,1)	12,4	0	0	10	38
Reading Comp. I	279	58%	2,3	(3,3)	5,5	0	0	5	13	276	57%	2,6	(3,6)	6,0	0	0	6	12
Productive Listening	279	4%	3,2	(1,4)	3,3	2	3	4	6	276	4%	3,5	(1,4)	3,6	3	4	4	6
Vocabulary	279	2%	5,1	(1,2)	5,2	5	5	6	6	276	2%	5,0	(1,4)	5,1	5	5	6	6
Phonemic Awareness	279	10%	3,3	(2,1)	3,7	2	3	5	10	276	7%	3,9	(2,0)	4,2	3	4	5	9
Expressive Vocabulary	279	0%	9,1	(3,5)	9,2	7	9	11	20	276	0%	9,6	(3,6)	9,6	7	9	12	20
Grade 2																		
Letters p/m	278	1%	44,8	(20,3)	45,4	34	49	59	93	283	1%	48,1	(20,1)	48,6	35	51	63	91
Di-/ Trigraphs p/m	278	15%	24,3	(19,6)	28,7	5	26	39	87	283	17%	27,1	(20,5)	32,6	6	30	43	80
CVCV Words p/m	278	20%	20,4	(16,4)	25,6	4	19	33	79	283	17%	22,6	(17,1)	27,2	7	22	35	67
Familiar Words p/m	278	22%	14,7	(11,9)	18,8	3	15	24	45	283	18%	16,8	(12,9)	20,4	5	18	27	54
Oral Reading Fluency	278	16%	16,7	(14,0)	19,9	4	16	26	63	283	17%	18,2	(14,8)	22,0	3	19	29	68
Reading Comp. I	278	18%	6,4	(4,0)	7,9	3	8	10	14	283	19%	6,7	(4,1)	8,3	3	8	10	14
Productive Listening	278	1%	3,9	(1,2)	4,0	3	4	5	6	283	1%	4,1	(1,2)	4,1	3	4	5	6
Vocabulary	278	0%	5,5	(0,8)	5,5	5	6	6	6	283	1%	5,4	(1,0)	5,5	5	6	6	6
Phonemic Awareness	277	2%	5,0	(1,9)	5,1	3	5	6	10	283	1%	5,1	(2,0)	5,2	4	5	6	10
Oral Reading Fluency II	278	24%	15,3	(13,5)	20,3	1	15	25	59	283	22%	16,3	(13,3)	20,9	3	16	26	60
Reading Comp. II	278	27%	4,0	(3,1)	5,5	0	5	6	10	283	23%	4,3	(3,2)	5,6	1	5	7	10
Sentence Choice	278	27%	5,4	(3,8)	7,3	0	7	9	10	283	29%	5,6	(4,0)	7,8	0	7	9	10

Figure A1: Percentage of copies lost or destroyed – Teacher report



12.1. Delivery And Use Of Vula Bula Anthologies In Funda Wande Impact Evaluation Schools

12.1.1. DATA SOURCES

These preliminary findings on the delivery and use of the Vula Bula anthologies are based on the following data sources:

- Interviews with the principal, deputy-principal or HOD in 57 schools
- Interviews with 61 Grade 1 and 63 Grade 2 teachers in 57 schools
- Interviews with 555 Grade 1 and 561 Grade 2 learners in 58 schools
- Audits of 54 Grade 1 and 54 Grade 2 classrooms in 55 schools

12.1.2. DELIVERY AND DISTRIBUTION OF VULA BULA ANTHOLOGIES

Almost all respondents (93%) to the principal/HOD interview reported that the Vula Bula anthologies were delivered to the school. ***Table A6*** shows the percentage of these schools that report sufficient copies of the anthologies being delivered for each learner to have their own copy.

Table A6: Percent of schools receiving sufficient anthologies for each learner – Principal report

	Grade 1 learners	Grade 2 learners	Grade 3 learners
Grade 1 anthology	94%	57%	
Grade 2 anthology		94%	57%
Grade 3 anthology			87%

Principals and HODs report almost universal (96%) distribution of anthologies to individual learners in classes.

The vast majority (96%) of teachers report receiving copies but 1% did not receive enough copies for each learner to have their own copy. Teachers were asked how many of the copies that they received at the beginning of the year had been lost or destroyed by term 4 ***(Figure A1***). Just over a third (34%) of teachers report that no copies have been lost and a further 34% report that almost no copies have been lost.

Ninety-one percent (91%) of learners report that they were given their own copy of the Vula Bula anthologies. Almost all (90%) of those who were given their own copy, still have the copy.

During the visit to the selected Grade 1 and Grade 2 classrooms, the

enumerators asked the learners to show them their copy of the anthologies. ***Figure A2*** shows the percentage of learners in each of the 108 classes that were able to show their copies. In one in five classes, 90% to 100% of learners had their own copy to show the enumerator. In a further 22% of classes, more than 75% of the learners had their own copy. In 5% of classes no learners were able to show a copy.

In classes where less than 75% of learners were able to show their own copies, the enumerators were instructed to ask the learners why they had no copy. ***Table A7*** shows the distribution of the main reasons given in the 59 classrooms where less than three-quarters of learners had copies with them. In 75% of these classes, books being at home was given as the main reason or one of the main reasons. Books being lost was reported as a reason in 35% of these classes.

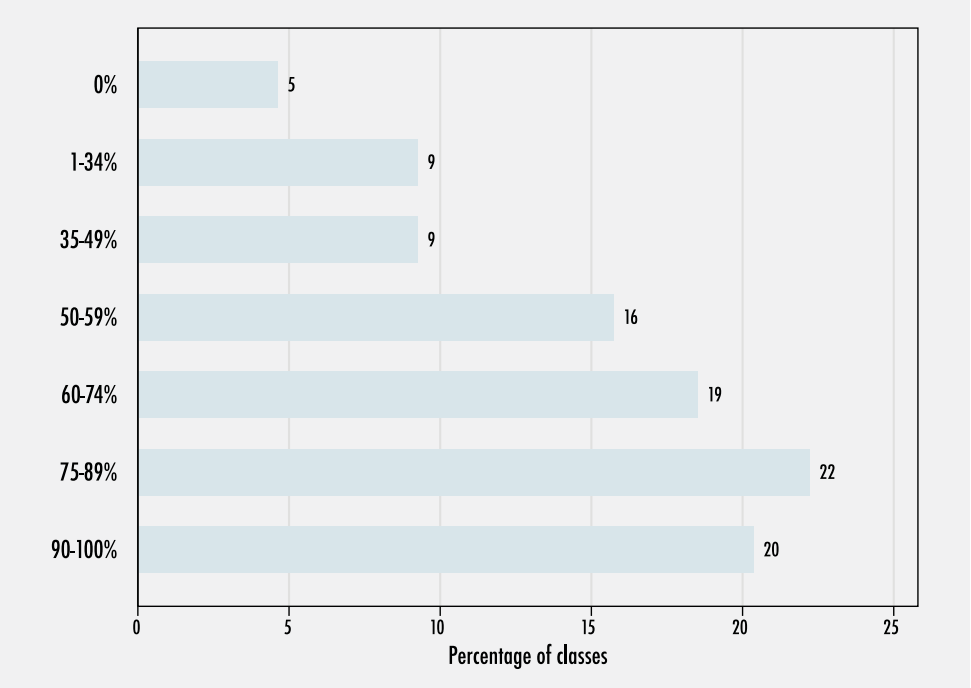
Table A7: Main reason for learners not being able to show books in classes where less than 75% of learners were able to show their own copy – Classroom observation

	Percent of classes
Lost	23%
At home	58%
Lost/At home	12%
At home/learner did not get books	5%
Not enough for every learner to have own copy	2%

12.2. Are Learners Allowed To Take The Vula Bula Anthologies Home?

We asked this question of principals/HODs, teachers and learners. In the schools that received anthologies, 96% of principals/HODs report that the learners are allowed to take the anthologies home. A similar percentage (97%) of teachers who received anthologies say they allow the learners to take the anthologies home. Ninety-two percent (92%) of learners report being allowed to take the Vula Bula anthologies home.

Figure A2: Percentage of learners able to show their own copy – Classroom audit



12.3. How Well Used Are The Vula Bula Anthologies?

Of the teachers who received copies of the anthologies, all but one report using them in class. ***Figure A3*** shows the frequency of reported use. Around two in five teachers report using the anthologies every day and 54% use them multiple times per week.

Table A8 shows the range of activities that the teachers conduct

Table A8: Anthology use in class – Teacher report

	Percent
Teacher reads aloud from anthology while learners listen	46%
Whole class reads from book at same time	22%
Small groups during group guided reading	69%
Shared reading/paired reading	68%
Independent reading during break/DEAR period	47%

Figure A3: Frequency of use of Vula Bula stories in class – Teacher report

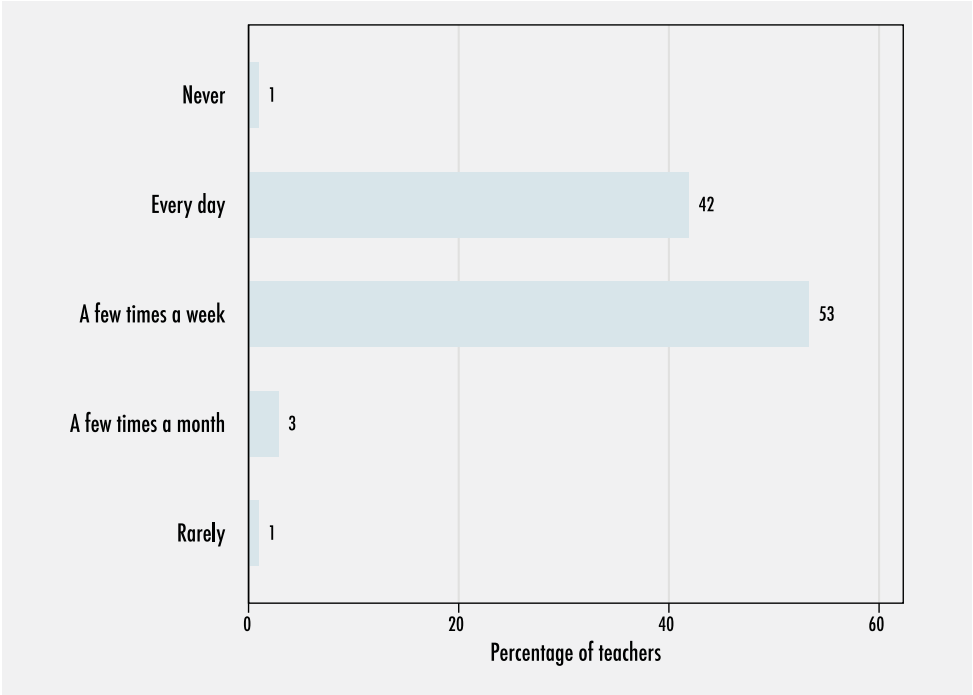
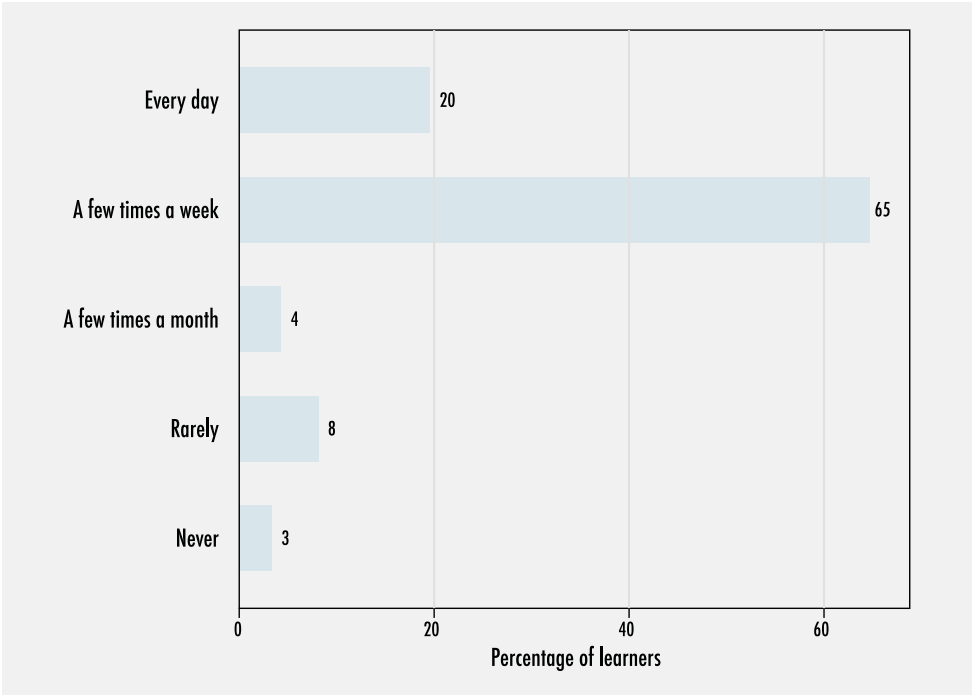


Figure A4: Frequency of use of Vula Bula stories in class – Learner report



with the anthologies. The anthologies appear to be used across a wide range of reading activities

Figure A4 shows the distribution of learner responses to the question about how frequently the Vula Bula readers are used in class. Twenty (20) percent of learners report that they are used daily and 65 percent report they are used multiple times per week.

Learners were also asked how often they read the Vula Bula anthologies at home. ***Figure A5*** shows that 85% of learners report reading them at least once per week.

In the Grade 2 classroom visit, the enumerators were asked to examine the anthologies and record how well used they appeared to be ***(Figure A6)***. In 68% of the classrooms, the anthologies were considered well used and in an additional 24% of classrooms they appeared to have been moderately used. Only 4% of classrooms had anthologies that were classified as so well used that they were dirty, torn or damaged.

The next two figures show the distribution of the number of stories that teachers and learners report using. ***Figure A7*** shows that 51% of teachers have used all or almost all of the stories.

Learners were asked how many of the stories they had read themselves or had read to them by their teacher or someone at home. ***Figure A8*** shows that 32% of learners report having read only a very few stories and 30% report having read all or most of them.

The teacher and learner reports, together with the classroom audit suggest extensive use the Vula Bula anthologies in class and at home. In an additional attempt to gauge use of the anthologies, the enumerators asked learners to show them their favourite story in the anthology they were most familiar with[54]. ***Figures A9 and A10*** show the distribution of favourite stories for the Grade 1 and Grade 2 anthologies respectively. Nearly three in ten learners using the Grade 1 anthology selected the first story as their favourite and a further 29% selected stories 2, 3 and 4.

Figure A5: Frequency of reading Vula Bula stories at home – Learner report

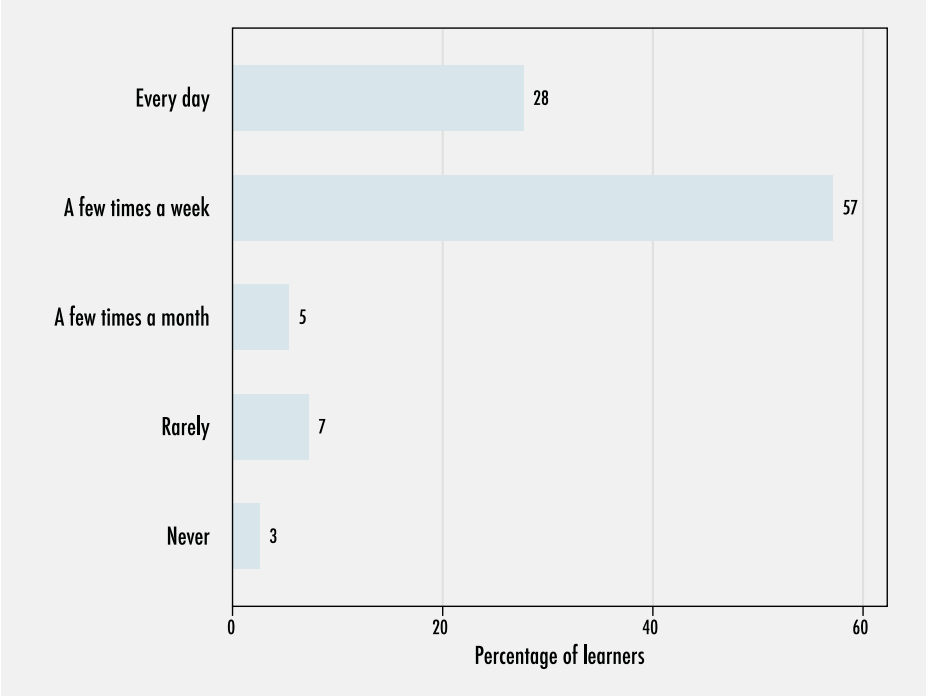
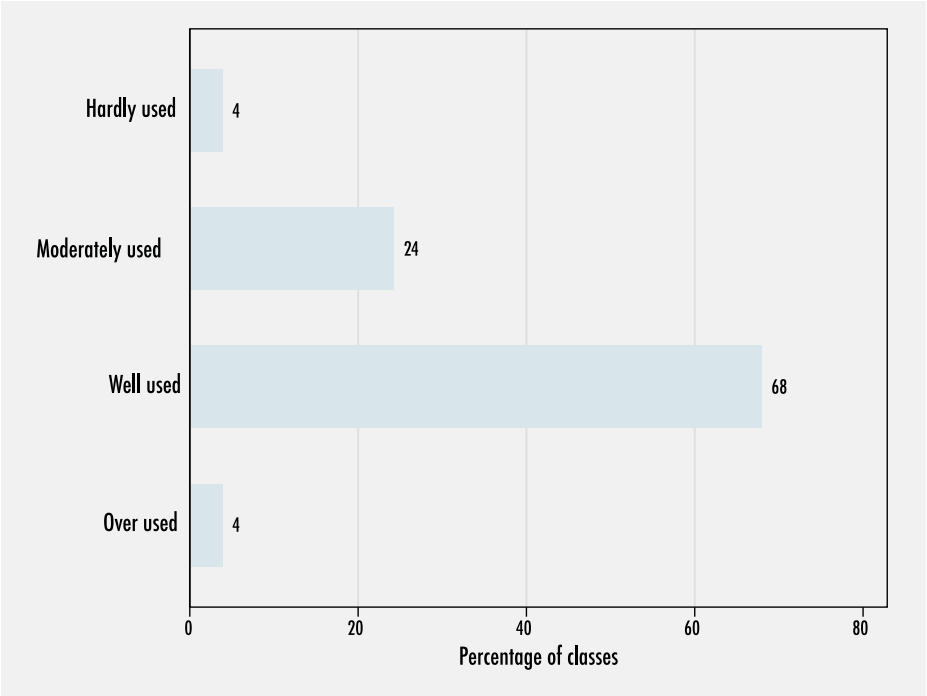


Figure A6: Condition of anthologies in Grade 2 classrooms – Classroom audit



One in five learners familiar with the Grade 2 anthology selected the first story as their favourite. Similar to the Grade 1 anthology, stories near the beginning tend to be most popular. This could be driven by learners favouring the easier stories or suggestive that some learners are less familiar with all the stories in the

54. Thirteen Grade 2 learners were most familiar with the Grade 1 anthology. All other learners were familiar with the anthology for their Grade.

Figure A7: Number of Vula Bula stories used this year – Teacher report

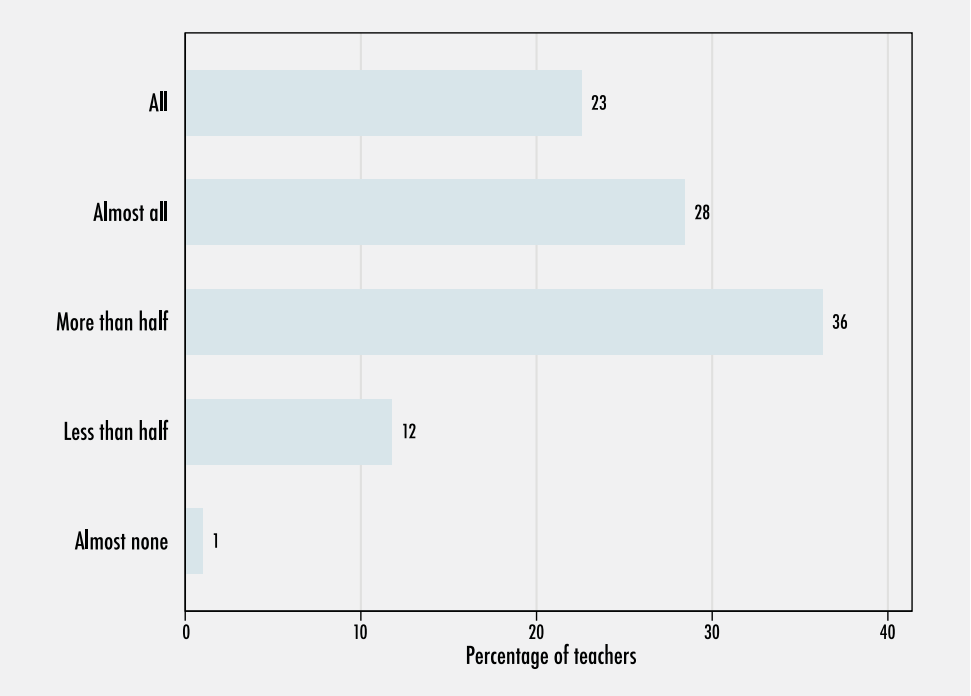


Figure A8: Number of Vula Bula stories read – Learner report

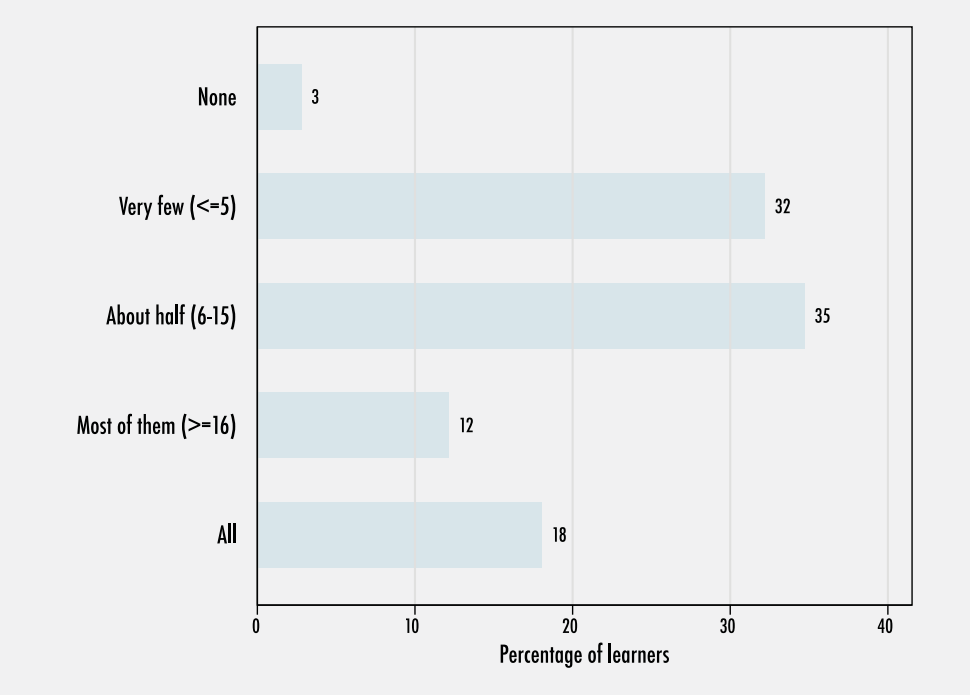


Figure A9: Favourite story in Grade 1 anthology – Learner report

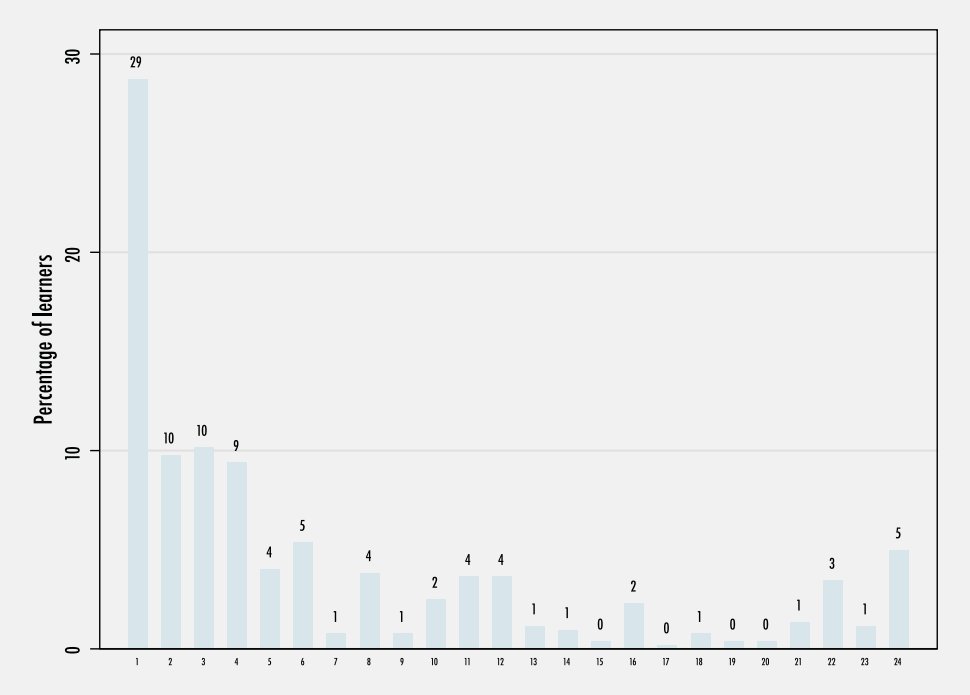
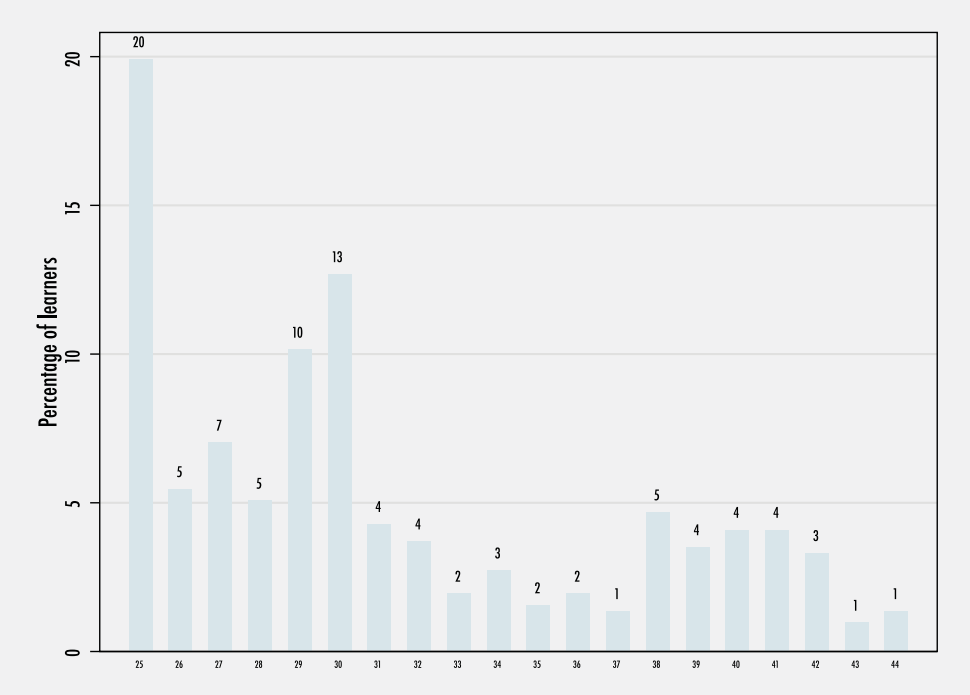


Figure A10: Favourite story in Grade 2 anthology – Learner report



anthology.

The enumerators also showed the learners a specific story in the anthology and asked if they had read that story or had somebody read it to them (the third story in the Grade 1 anthology and the second story in the Grade 2 anthology). If they replied yes, the enumerator asked a simple question about the story. Ninety percent of learners familiar with the Grade 1 anthology said they had read the story and 86% of those learners were able to correctly answer the question. The analogous figures for the Grade 2 anthology were 85% and 46%.

12.4. Are The Vula Bula Anthologies Grade Appropriate?

Teachers were asked whether they felt the level of the stories were appropriate for their Grade. The vast majority (91%) of teachers felt that the stories are at the right level, while 8% felt that they were too easy. Very few teachers felt that the stories were too difficult.

An example page from each of the three anthologies was shown to the teachers and they were asked how many of their learners would be able to read the text on their own. The responses are shown in ***Table A9***. On average, Grade 1 teachers felt that 60% of their learners could read the Grade 1 text but only 48% and 37% would be able to read the Grade 2 and Grade 3 texts respectively. A similar percentage (64%) of Grade 2 teachers thought their learners could read the Grade 2 text. On average, Grade 2 teachers feel that around one quarter of their learners would not be able to read the Grade 1 level text on their own. Grade 2 teachers believe that one in two of their learners would manage to read the Grade 3 text on their own.

Table A9: Percentage of learners in class able to read text at this level – Teacher report

	Grade 1 teacher	Grade 2 teacher
Grade 1 anthology	60%	74%
Grade 2 anthology	48%	64%
Grade 3 anthology	37%	50%

12.5. Future Provision And Training

Just over three quarters (76%) of teachers believe the ECDOE will provide these anthologies for every learner every year and 35% of the teachers had attended the anthology training

Figure A11: Grade appropriateness of Vula Bula anthologies – Teacher report