IMPACT EVALUATION OF FUNDA WANDE IN-SERVICE TEACHER COACHING INTERVENTION

FINDINGS FROM THE FIRST YEAR

CHRISTIAAN MEIRING

MRNCHR004

JUNE 2020

A dissertation submitted in partial fulfilment of the requirements of:

Master of Commerce in Applied Economics

Plagiarism Declaration

COMPULSORY DECLARATION:

1. This dissertation has been submitted to Turnitin (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.
2. I certify that I have received Ethics approval (if applicable) from the Commerce Ethics Committee.
3. This work has not been previously submitted in whole, or in part, for the award of any degree in this or any other university. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works of other people has been attributed, and has been cited and referenced.

Student number	MRNCHR004
Student name	Christiaan Meiring
Signature of Student	Signed by candidate
Date:	26 June 2020

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to Professor Cally Ardington. I have learnt lasting lessons from her guidance and supervision, not only on the technical aspects related to the dissertation, but also on how to keep one's eye on what is truly important. I've been extremely fortunate to work with someone whose expertise and passion for her field is matched by an equal measure of mentoring ability and patience (and a willingness to save me from myself when necessary).

My thanks to Professor Murray Leibbrandt for his co-supervision and thoughtful comments. This dissertation is the culmination of a process which started with encounters a few years earlier and inspiration taken from his work, which drew me to the field in the first instance.

I would also like to acknowledge that this research benefited greatly from an ecosystem in SALDRU, where young researchers have access to opportunities, resources and the network of incredible people to succeed and contribute toward topics they care deeply about.

Financial support from the NRF National Research Chair of Poverty and Inequality Research, an opportunity to provide research assistance within SALDRU, and the UNU-WIDER's SA-TIED Young Scholars programme have made this research possible in a very practical sense - which is gratefully acknowledged.

I wish to express my deepest gratitude to my family for their loving support throughout my years of study: to my mother Barbara for always backing me and for unfailing encouragement; to my farther Ian whose sacrifice has afforded his children the opportunity to chart their own path.

Finally, this research is in dedication to my loving wife Nadia, who has been a rock and a source of strength throughout, and who's career as a primary school teacher herself has proved to be the most insightful source for insight and understanding.

CONTENTS

Student number 1

Student name 1

Signature of Student 1

Date: 1

Acknowledgements 2

Contents 3

Executive Summary 5

1 Introduction 7

2 Background and context 9

2.1 The Eastern Cape: Emblematic of low resource, low capacity schooling environments across South Africa 9

2.2 The lack of teacher capacity in low resource South African schools 10

3 Programme description 12

3.1 Funda Wande 12

3.2 Theoretical framework of the Funda Wande intervention 13

3.2.1 The state of evidence on improving learning outcomes in developing countries 13

3.2.2 The theory of how children learn to read for meaning 15

3.2.3 Shifting teacher instructional practice 20

4 Evaluation design 22

4.1 Research questions 22

4.2 Evaluation methodology overview 22

4.3 Recruitment and random assignment 23

4.4 Instruments 24

4.5 Baseline sample description 25

4.6 Midline sample, attrition and balance 29

5 Midline results 33

5.1 Descriptive statistics 33

5.1.1 Development of reading skills in the control group 33

5.1.2 Primary outcome measures 36

5.1.3 Descriptive analysis of intervention impacts 38

5.2 Estimating treatment effects 44

5.2.1 Regression analysis of intervention impacts 44

5.2.2 Intervention impacts 45

5.3 Heterogenous treatment effects 52
6 Intermediate outcomes: Teachers 62
6.1 Formative assessment and awareness of learners' actual reading levels 62
6.2 Use of graded readers 66
7 Discussion 69
8 References 74
9 Appendix 81
9.1 Appendix tables and figures 81
9.2 Annex I: Summary of the EGRS interventions and Results Comparison 86
9.2.1 What we learnt from EGRS I and II results 86
9.2.2 In depth comparison of Funda Wande and EGRS midline results 87
9.3 Annex II: Funda Wande programme inputs 89
9.4 Annex III: School selection and underlying population 97
9.5 Annex IV: Learner assessment details and rationale 99
9.6 Annex V: Intermediate outcomes – Teacher Formative Assessment (further analysis) 104

EXECUTIVE SUMMARY

Acknowledging the limited opportunities for South African teachers to acquire specialized knowledge in teaching reading, particularly in African Languages, the Non-Governmental Organisation Funda Wande is providing an integrated package of support to train Foundation Phase literacy teachers in how to teach reading for meaning in African languages. The pilot intervention evaluated here takes place in three urban schooling districts in South Africa's Eastern Cape province. The intervention is implemented in partnership with the Eastern Cape Department of Education (ECDoE). The Funda Wande in-service training model builds on international best practice and lessons learnt from domestic iterations of integrated teacher training and support models, prominently amongst which are the Department of Basic Education (DBE)-led Early Grade Reading (EGRS) studies

Funda Wande provides a bouquet of home language print resources to learners and classrooms. Teachers also receive an integrated package of curriculum aligned lesson plans, assessment booklets and online pedagogic resources. In-classroom teacher coaching provides support, monitoring and feedback for teachers on how to implement new teaching techniques and make use of materials. The Funda Wande course trains teachers and their Heads of Department (HODs) on how to teach reading for meaning in African languages. The course content provides teachers with knowledge on the morphology of African languages and how learners learn to read in them, whilst supporting materials and in-person coaching equips teachers to implement the instructional techniques in practice.

The primary aim of the evaluation is to assess whether the Funda Wande intervention is effective in changing teacher's instructional behaviour and improving early reading outcomes. More specifically, the programme's is evaluated with reference to its self-stated goal: that all learners should be able to read for meaning in their home language by the end of Grade 3. The specific reading outcomes assessed are grade relevant foundational reading and reading comprehension skills.

A randomized control trial (RCT) is used to estimate the causal impact of the programme on early literacy outcomes with schools randomized into one of two arms – Funda Wande and control – in three urban and peri-urban education districts. All schools in the evaluation are no fee public schools with an isiXhosa language of learning and teaching. After one year of programme exposure, the intervention impact on the treated group of Grade 1 and 2 learners' reading proficiency is 0.17 standard deviations. Translated into the amount of learning that took place in comparison schools, or 'business as usual' learning environments, this effect equals 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 comparison status quo classrooms.

The programme effects are positive across all the sub-domains of reading proficiency that could be measured reliably. 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 progress to reading for meaning. The impacts on downstream higher order reading comprehension skills are only detectable for Grade 2 learners. Consistent with other results from the recent literature, these findings support the idea that learners require a range of foundational literacy abilities before they can read with some level of fluency. In

turn, learners need to read with a certain minimum level of speed and accuracy 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 do 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. A related finding is suggestive evidence that the programme helps boys in treatment schools catch up with their generally more proficient girl counterparts, but only in Grade 2 and with the extent of catch-up 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 graded readers - which were provided equivalently to both treatment and control schools - more frequently; and (c) to use instructional techniques that have previously shown to facilitate more individualised forms of learner reading practice and -teacher feedback.

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.

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. Consistent with the results from similar interventions in Kenya, Uganda and South Africa, the Funda Wande intervention improves 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.

1 INTRODUCTION

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).

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 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 learner 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] (Motala and Carel, 2019; World Bank, 2018b).

In turn, poor learning outcomes in aggregate are also disproportionately driven by the majority of learners who find themselves in the poorest 75 to 90 percent of schools in the country's bimodal education system (Spaull and Pretorius, 2019; Spaull, 2013). 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 greatest potential for long run returns arguably lie in targeting interventions at learners early in their schooling careers. Core skills (like basic literacy and numeracy) are hierarchical, forming the foundations on which subsequent learning and skills development takes place (Heckman, 2006). When a learner falls behind, the deficit is compounded over time as there is a mismatch between the level of classroom instruction and learner's actual learning levels (Glewwe and Muralidharan, 2016). If learners do not master reading as a core skill, specifically, they cannot acquire further subject-specific

[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 in South Africa are not attending school (World Bank, 2018a).

[3] Mother tongue education is a Constitutionally mandated right for learners in Grade 1 to 3. Therefore, more than 70 percent of South African children learn to read in an African language before switching to either English or Afrikaans in Grade 4 (Pretorius and Spaull, 2016). The majority of Grade 4 children (approximately 90 percent) transition to English as language of instruction in Grade 4, with the remaining 10 percent generally receiving instruction in Afrikaans.

[4] South Africa's public expenditure on education has consistently been 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), as well as the OECD average of 5.2 percent (IMF, 2019).

knowledge that relies on the foundation of reading comprehension (Spaull and Pretorius, 2019; Spaull and Kotze, 2015).

A promising avenue for early intervention is through targeting teacher capacity specifically. 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 Rivkin, 2010; Chetty et al., 2014) and developing countries (Bau and Das, 2019; Bold, et al., 2017; Bruns and Luque, 2014). The focus on improving teacher 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[5]. 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).

The focus on improving teacher quality has seen an increase in programmes aiming to strengthen teacher capacity through professional development-, or structured pedagogical programmes. These programmes characteristically consist of:

i) an integrated package approach that includes the provision of curriculum aligned learning materials[6] (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 consisting of initial teacher training, implementation support, feedback and/or mentoring) (Cilliers et al., 2019, Fleisch et al. 2016).

Acknowledging the limited opportunities for South African teachers to acquire specialized knowledge in teaching reading, particularly in African Languages, the non-governmental Organisation Funda Wande has designed a course and integrated package of support to train Foundation Phase literacy teachers in how to teach reading for meaning in African languages. The programme design builds on insights from domestic and continental iterations of similar structured pedagogical interventions and teacher professional development programmes.

The Funda Wande intervention provides a carefully designed bouquet of home language print resources to learners and classrooms. Teachers also receive an integrated package of curriculum aligned lesson plans, assessment booklets and online pedagogic resources. Besides the innovative design of these specific programme components, what differentiates the Funda Wande programme from previous structured pedagogic interventions in the domestic context is its particular emphasis on coupling knowledge of the process by which children learn to read (in terms of the core domains of decoding, reading comprehension and learner affective response) with the "linguistic and orthographic underpinnings of early reading instruction" in different African languages (Funda Wande, 2018). The intervention aim is to equip teachers with both deep foundational knowledge and the necessary resources as part of the broader pedagogical strategy.

[5] Notably the range of pilot (PRIMR) and at scale (Tusome) studies by Piper, Zuilkowski and colleagues in Kenya (see Piper et al., 2014, 2015, 2018b and Zuilkowski and Piper, 2017). See also similar programmes in Liberia (Piper and Korda, 2011) and Uganda (Kerwin and Thornton, 2019)

[6] Often referred to as Learning and Teaching support materials (LTSM).

Partnering with the Eastern Cape Department of Education (ECDoE), Funda Wande is implementing a pilot structured pedagogical intervention delivered by means of in-service teacher coaching in selected schools from urban and peri-urban areas in three education districts in the Eastern Cape. All schools are no fee public schools with an isiXhosa language of learning and teaching. This paper evaluates the impact of the intervention after the first academic year of implementation. The primary aim of the evaluation is to assess whether the intervention is effective in changing teacher's instructional behaviour and subsequently improving home language early reading outcomes. A randomized control trial (RCT) design is used to estimate the causal impact of the programme on early literacy outcomes.

The paper proceeds as follows: section 2 provides background to the study and describes the context facing teachers in low resource public schools across South Africa and in the Eastern Cape specifically, section 3 provides a description of the Funda Wanda intervention, section 4 describes the evaluation design, results are reported in section 5, and section 6 concludes with a discussion of how findings here relate to the larger literature and future research.

2 BACKGROUND AND CONTEXT

2.1 THE EASTERN CAPE: EMBLEMATIC OF LOW RESOURCE, LOW CAPACITY SCHOOLING ENVIRONMENTS ACROSS SOUTH AFRICA

The Funda Wande intervention specifically targets no fee, public schools where the foundation phase language of learning and teaching (LOLT) is an African language. These schools represent a significant share of the South African schooling system. The South African education system consists of both government (public) and private (independent) schools. At the primary school level specifically, about 93 percent of learners attend public schools (Howie et al., 2017). Public schools can be further divided into the relatively well-functioning, fee paying section and the generally low resource, low capacity, no fee schools. Approximately 87 percent of South African public schools are non-fee-paying schools and they educate more than 70 percent of the learners in the country (Howie et al., 2017).

The socioeconomic status (fee-paying status) of the school that a learner attends is the single greatest predictor of their academic performance (Taylor, 2011). In the majority of cases learners are only able to attend schools near to where they live – which in the South African context implies that learners from low resource communities do not generally have the option to attend schools of reasonable quality. However, given the country's historical legacy of racial segregation, a school's socioeconomic status is also highly correlated with its geographic location (both province and urban/rural setting), race and the language of instruction in the early grades (Spaull, 2019).

Based on the latest international testing data[7] (Ishdale et al., 2017), isiXhosa is the home language of approximately a fifth of the country's Grade 5 learners - making it the second most prevalent home language (after isiZulu). English, the language of instruction for approximately 90 percent of leaners from Grade 4 onwards, is the home languages of less than ten percent of learners. Learners from African language backgrounds are therefore required to become both bilingual and biliterate from Grade 4 onwards. Illustratively, only 12 percent of isiXhosa learners could reach the lowest benchmark

[7] The Trends in International Mathematics and Science Study (TIMMS), multi-country standardised tests.

for literacy proficiency in the 2016 PIRLS was, compared to an already low 43 percent of English home language learners (Howie et al., 2017)

It is not yet clear to what extent the differential performance of learners in isiXhosa is attributable to the inherent difficulty of learning to read in the language itself. As Spaull (2016) shows, learners' home language and the LOLT of their school are strongly correlated with other factors that predict academic performance. isiXhosa classrooms are more likely to be in less well-resourced and in less capable schools, -communities and -administrative districts. For example, from the 2016 PILRS data, the average Grade 4 class size was 46 for learners who completed the assessment in isiXhosa, compared to the 35 and 42 learners in the average Afrikaans and English classrooms respectively.

Focussing on the Eastern Cape province specifically, the LOLT for foundation phase learners in Grades 1 to 3 is predominantly either English or isiXhosa. The home-language background for learners within the same schools in the province is also fairly homogenous, when compared to the rest of the country (Spaull and Pretorius, 2019). Overall, 82 percent of the Grade 4 learners in the Eastern Cape wrote the 2016 PIRLS literacy assessments in their home language, compared to 71 percent in the country as a whole (and as low as 40 percent in Gauteng).

For learners to learn how to read, a necessary condition is that they have access to reading materials in the language that they are learning to read in. However, only 30 percent of schools have access to some form of library (DBE, 2018). This is compounded by the fact that the libraries that do exists are often poorly stocked (especially with regards to African language reading materials). Consistent with the general theme , learners in formerly "white only", fee-paying primary schools are far more likely to have access to libraries at school (87 percent of learners) than learners in schools formerly classified as rural homelands[8] or urban African schools (where only 35 - 36 percent of learners today have access to libraries at school, in any form) (DBE, 2014).

2.2 The lack of teacher capacity in low resource South African schools

Besides the challenge of large and heterogeneous classes, the majority of South African teachers who work in low-resource, no-fee public schools have a lack of educational resources, they are generally not equipped nor supported to effectively manage these resources, and they have historically not received any meaningful training in how to teach reading (Taylor and Taylor, 2013; Van der Berg et al., 2016). The result is an inequality-exacerbating scenario where the learners with the largest early learning deficits generally find themselves in the schools with the least capacity (Altinok, 2013, Gustafsson, 2016, NEEDU, 2013; Spaull, 2015; Spaull and Pretorius, 2019, Venkat and Spaull, 2015).

Teacher content- and pedagogical knowledge is a binding constraint on improving the performance in the majority of low-resource South African schools. This despite South African teachers generally being adequately qualified with respect to academic qualifications (Taylor et al., 2013) and being on the receiving end of a plethora of (often competing) supporting strategies and models[9] (Van der Berg

[8] Homelands were areas established by the Apartheid government, to which the majority of citizens classified as "non-white" were forcibly moved with the aim of separating people of different race categories.

[9] Worth noting, however, is that until very recently there has been little to no rigorous evaluation of these competing teacher- and principal support and professional development programmes, leaving policymakers and researchers with little sense of which programmes are working, and why they are (not) working (Van der Berg et al., 2016).

et al., 2016). This lack of preparation has not been effectively addressed in in-service professional development (Shalem and De Clerc, 2019, Taylor, 2019) or tertiary educational training (Initial Teacher Education Research Project, 2014; Taylor, 2019).

Teaching is a complex task that requires teachers to understand and operationalise knowledge from diverse domains simultaneously (Taylor, 2019). Teachers' subject matter mastery and pedagogical knowledge are perquisites for being effective teachers (see Carnoy and Arends, 2012; Taylor, 2014; Taylor and Taylor, 2013). Teachers must have the *disciplinary knowledge* of the subjects that they teach[10] and they must also understand the process by which reading skills are acquired, how this differs by language structure, as well as the best strategies and methods of teaching reading in the specific language (*pedagogical content knowledge*). This entails understanding in a very exact manner how learners learn to read and how to intervene at different stages in learners' developmental trajectories (Taylor, 2019). Teachers must also understand how the specific skills and subject matter that they teach fits within the broader curriculum (both to the other subjects and later grades - *curricular knowledge*). On top of this, teachers must translate these diverse knowledge components into a coherent and effective sequence of classroom activities and practices (*pedagogical competence*) (Hoadley, 2016, Van der Berg et al., 2016).

In practice, common findings from classroom observations on norms in teacher instructional practices suggest that these do not align with best practice teaching methods, lesson pacing or curriculum coverage (Hoadley 2012, 2016, 2018; NEEDU 2013; Taylor and Taylor 2013). Illustratively, often the principal method for teaching reading is to read to the class and have learners collectively chorus back what was read (e.g., Prinsloo 2008). Conversely, there is often a lack of individualized learner attention (NEEDU 2013) and a neglect of formal instruction of other skills foundational to learning to read (such as phonics, spelling, and vocabulary; see Spaull and Pretorius 2019).

In an environment where both teacher absenteeism (Reddy et al. 2010; Spaull 2011) and curriculum coverage are low[11], current evidence suggests that it is the lack of productive teaching activity taking places *despite* the presence of teachers that is a major binding constraint to learning (Carnoy et al. 2012; Hoadley and Gallant 2016; Van der Berg et al. 2016). In terms of the causes of learners' lack of opportunities to learn, there are two explanatory factors for which there is evidence at the national, systemic level (Van der Berg et al., 2016, Spaull, 2019:): i) the lack of teacher capacity (in terms of subject content knowledge, pedagogical competence, etc) and ii) a lack of accountability[12] (in terms of monitoring and support for teachers by senior school staff, principals, and district officials). Both factors arguably play a complementary role in constraining learning outcomes in South Africa (Van der

[10] In contrast, a common finding in South African primary schools is that teachers of lack the basic disciplinary knowledge at the level of the learners that they are expected to teach (Taylor and Taylor, 2013; Van der Berg et al., 2016; Venkat and Spaull, 2015).

[11] Evidenced in various systemic evaluations: the National School Effectiveness Study, National Education Evaluation and Development Unit case studies, and the School Monitoring Survey (see Dechaisemartin 2013; NEEDU 2013; DBE, 2015).

[12] In depth accounts on state of teacher accountability structures in the South African education system are provided by Van der Berg et al (2016) and a Ministerial Task Team report under the stewardship of Professor John Volmink (DBE, 2016).

Berg et al., 2016). Evidence on teacher coaching programmes suggests a way forward, by improving teacher capacity whilst also building relationships of professional accountability.

3 PROGRAMME DESCRIPTION

3.1 FUNDA WANDE

The Funda Wande intervention builds upon the lessons learnt from the international literature and promising insights from previous iterations of similar approaches to improving teacher instruction and learner learning outcomes in the South African context - most notably the Department of Basic Education-led Early Grade Reading (EGRS) studies (Cilliers et al., 2019; Kotze et al., 2019)[13].

For these packaged interventions (often categorised as structured pedagogical interventions – Sniltsveit et al., 2016) to be successful, indications are that some degree of teacher support, monitoring and feedback are required. Furthermore, teacher training/professional development that focuses on specific instructional techniques, lesson planning, the effective use of complementary provided materials; as well as the implementation of more technically demanding teaching techniques (such as group-guided reading) are all contributing constitutive components.

Within structured pedagogy programmes, the mode of delivery has played an important role in programme effectiveness (Popova et al., 2018)[14]. For example, on-site teacher coaching (as opposed to centralised training workshops) has proved to be an especially important component. Evidence for this comes from both meta-analyses of international evidence (Kraft et al, 2018) and domestic iterations comparing the relative effectiveness of coaching versus centralised training (Cilliers et al., 2019). The idea behind in-classroom coaches is often to provide teachers with the support, monitoring and feedback required to integrate new materials and novel pedagogical techniques into their daily classroom practices. 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"*

Drawing on these insights, the Funda Wande programme makes use of a carefully designed[15], multi-media course to train Foundation Phase (Gr R-3) teachers (using professionally filmed in-classroom videos, info-graphics and other multi-media). The course teaches the major components of reading

[13] Earlier examples include the Gauteng Primary Language and Mathematics Strategy programme (Fleisch et al., 2016; Fleisch and Schoër, 2014), the Reading Catch-Up Study (Fleisch et al., 2017), the Systematic Method for Reading Success study (Piper, 2009), and the Learning for Living project (Sailors et al., 2010).

[14] 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).

[15] The Funda Wande literacy course and materials were developed over two years with input from over 15 South African academics from five universities. The course is nationally (SAQA) accredited and has strong support from the national Department of Basic Education, the Eastern Cape Department of Education and Rhodes University.

and writing in isiXhosa (the pilot language), with subtitles in English. The essential components of the 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 learners. 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. 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's Curriculum and Assessment Policy Statements (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 after school (three per term), and one-on-one in-classroom visits with each teacher in Foundation Phase (at least once per term).

Head of department (HOD) training: 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).

3.2 THEORETICAL FRAMEWORK OF THE FUNDA WANDE INTERVENTION

The Funda Wande intervention package targets the poorly functioning "instructional core" in most no fee, public schools - the "actual interaction between teachers, learners, and content in the classroom" (City et al., 2009). The composition of the intervention is based on insights gained from three somewhat disparate strands of literature.

3.2.1 THE STATE OF EVIDENCE ON IMPROVING LEARNING OUTCOMES IN DEVELOPING COUNTRIES

Within the education production function framework (Glewwe et al, 2014), economists often think of schools as institutions that produce student learning (the 'output' it wants to maximise) by transforming a range of inputs through some technology. The inputs include books, physical school infrastructure, teachers, school management, parent involvement, etc., and the technology refers to the way these inputs together are brought to 'produce' learning.

The earlier experimental literature focussed on providing different kinds of seemingly lacking resources to schools, teachers and households in resource constrained environments (Kremer et al., 2013 provide an overview). Findings suggested that simply supplying inputs on their own, without

complementary guidance, incentives and/or accountability structures are not enough to shift learning outcomes[16]. Neither flipcharts nor textbooks[17] in Kenya (Glewwe et al., 2004, 2009), nor flexible grants in Niger (Beasley and Huillery, 2017), nor libraries in India (Borkum et al, 2013), nor anticipated grants in Zambia and India (Das et al., 2013) had any impact on test scores. The same goes for the reduction in class sizes through the provision of extra contract teachers in Kenya (Duflo et al, 2014) and giving schools extra computers in Columbia (Barrera-Osorio and Linden, 2009) or Peru (Cristia et al, 2012).

The main lesson from earlier studies is that business as usual input provision is rarely effective, and often expensive (Glewwe and Muralidharan, 2016). Research on South African schools also similarly suggests that the provision of additional school resources often had no impact on learning outcomes, because they were not well managed by the schools (Van der Berg, 2008, Taylor, 2013). Subsequent programmes have focussed on providing a combination of physical inputs (like textbooks, Information and Computer Technologies and teacher guides with lesson plans) integrated alongside complementary intervention components like teacher professional development, coaching, community interventions and personalised computer assisted learning programs (Piper et al., 2018).

The rapid growth in the number of experimental and quasi-experimental studies[18] has given rise to multiple narrative and systematic reviews of "what works" in improving learning outcomes, each providing an array of conclusions on what has proven successful in developing country contexts (for e.g. Conn, 2017; Glewwe and Muralidharan, 2016, Ganimian and Murname, 2016, McEwan, 2015, Snilstveit et al. 2016). Yet these reviews often differ on the exact composition of studies which they consider (based on inclusion criteria) and their categorisation of different intervention types based on i) different levels of disaggregation of intervention types[19] and ii) by focusing on or emphasising different elements of interventions (with interventions themselves often multi-pronged and part of complementary "bouquets"). The subsequent recommendations are somewhat different in foci and often provide policy suggestions that are not easily reconcilable.

Nevertheless, in an attempt to synthesise the systemic reviews of the high-quality empirical evidence on the interventions[20] that have improved learning outcomes in developing countries, Evans and Popova (2016) highlight two classes of programmes that have shown positive effects with some consistency:

> i) pedagogical intervention that tailor teaching to learner's actual learning levels (instead of the rigid expected levels of curriculums), either by means of teacher methodology or adaptive learning software; and

[16] Note that "learning outcomes" as used within this literature is often used with reference to test scores. This does not imply that other outcomes are not central to the educational process, but test scores are generally used as measurable proxies for underlying literacy, numeracy and other subject specific content knowledge.
[17] Relevant to note for the importance of personalised, right-level learning is that textbooks in Glewwe et al. (2009) did increase test scores for one subset of learners: the highest achieving students.
[18] High quality evidence generally refers to research that attempts to identify causal effects by establishing a well-defined counterfactual (with somewhat different cut-off points in what is considered convincing evidence on the spectrum between quasi- and RCT experimental research).
[19] For example, reviews can consider teacher training and teacher coaching separately, or teacher professional development as an overarching category.
[20] Note that "interventions" and "programmes" are used interchangeably here, with reference to the overall package intended to shift learning outcomes that is being evaluated.

ii) individualised, repeated teacher training/coaching interventions that promote a specific task or tool.

3.2.2 THE THEORY OF HOW CHILDREN LEARN TO READ FOR MEANING

The intermediate goal of structured pedagogy interventions more generally is to improve the content and quality of instruction, often including changes to the teacher's pedagogical methods (Sniltsveit et al., 2016). However, the ultimate outcome of interest is improvements in learners specific learning domains, such as reading proficiency. In the case of the Funda Wande intervention, the outcome of interest is that all learners should be able to read for meaning. Any shift in teaching techniques and classroom practices only serve as a means to this end.

3.2.2.1 THE ROLE OF MOTHER TONGUE

International evidence suggests that school entering children are best equipped to learn to read in the language that they are immersed in on a daily basis – their mother tongue (Ball, 2010). In turn, these learners are able to use foundational literacy skills acquired in their home language, such as how to decode words, to better position them to learn subsequent languages[21]. Learners might know how to decode in their home language and thus have a reference framework for figuring out letter-sound relationships in the new language, but their word recognition ability, vocabulary knowledge, and broader knowledge of the new language's differing characteristics (in terms of its orthography, morphology and grammatical structures) still need to be acquired (Goldenberg, 2013).

The importance of learning to read in one' home-language first reading is echoed in the limited evidence from the South African context. Taylor and Von Fintel (2016) find that learners who receive home language instruction in the first three years of schooling also fare better in subsequent English acquisition in Grades 4, 5 and 6. Similarly, the coaching intervention from the home-language targeted EGRS I study also led to a significant improvement in learners' first additional language (English) reading proficiency (Taylor et al., 2017)[22].

Even though the effectiveness of structured pedagogical programmes in improving early grade reading is well understood, the role of the specific LOLT used is not, neither in South Africa nor elsewhere on the continent (Brunette et al., 2019) [23]. Just as the impacts of structured pedagogy interventions are generally larger on language subjects than mathematics (Sniltsveit et al., 2016), we might expect the effectiveness of these programmes to also differ by the characteristics and complexity of the specific language itself. More concretely, early reading pedagogical practices in South African schools are

[21] This is referred to as the interdependence hypothesis (Cummins 2001, 2007). See also the discussion in Piper et al. (2016) for a recent review of the state of evidence.

[22] Another channel in the EGRS I intervention can also explain the positive second language effects: improved teacher ability. Teachers could plausibly have transferred their improved pedagogical techniques to their English teaching practice, thereby improving learners' outcomes in a language subject not specifically targeted by the intervention.

[23] In an exceptional case where authors have evaluated the heterogeneity of a reading programme's impacts based on language characteristics in Uganda (Brunette et al., 2019), they find large variations in the same programme's effectiveness depending on the complexity of the language in which the programme had to be delivered. In turn, implementing language characteristics also proved more predictive of the eventual programme impacts than either measures of programme implementation fidelity or learner socioeconomic status.

generally derived from teaching English (Funda Wande, 2018) and therefore not adapted to language specific characteristics[24].

Two differences in the language structures of English and isiXhosa illustrate the point. First, the orthography (or spelling system) of the English language is opaque, whereas it is transparent in isiXhosa. Therefore, in English the letters of the alphabet have a one-to-many relationship with the sounds that they represent[25], where it is a one-to-one relationship in isiXhosa (Funda Wande, 2018; Spaull et al, 2020). Second, English is an analytic language, whereas isiXhosa is an agglutinating language. This means that isiXhosa sentences generally consist of fewer words of longer length (because affixes are attached to word stems to change the meaning of words and/or sentences) (Funda Wande, 2018; Spaull et al, 2020)[26].

This differentiates the Funda Wande programme from previous structured pedagogic interventions: its focus on the "linguistic and orthographic underpinnings of early reading instruction" (Funda Wande, 2018: 11).

3.2.2.2 THE READING ACQUISITION PROCESS

The Funda Wande programme delineates three main components of the reading acquisition process, which in turn guides its programmatic design: i) decoding, ii) comprehension and iii) learners' affective response to text. **Decoding** consists of a range of skills necessary for deciphering unfamiliar words or word parts and translating written symbols into language. It is the first step and comprises the foundational skills on which further reading skills are built. Unlike oral language skills, decoding is usually not acquired naturally by young children over time. Instead, it must be taught systematically and continuously practised, with learners requiring repeated, individualised teacher feedback on their progression[27]. The decoding component of the intervention provides an understanding of the role of four subcomponents required for decoding (Funda Wande, 2018):

Phonological- and phonemic awareness: being aware of sounds in language and being able to hear, recognise and manipulate the specific sounds within words, respectively.

Alphabetic knowledge and phonics: understanding how letters and letter combinations on a page translates to sounds and vice-versa.

Word recognition: the ability to rapidly and automatically recognise a written word (or parts thereof), without having to first sound out the constituent letters. In turn, this frees up attention and memory for the other necessary tasks in reading for meaning (like focussing on the reading the next words in a sentence).

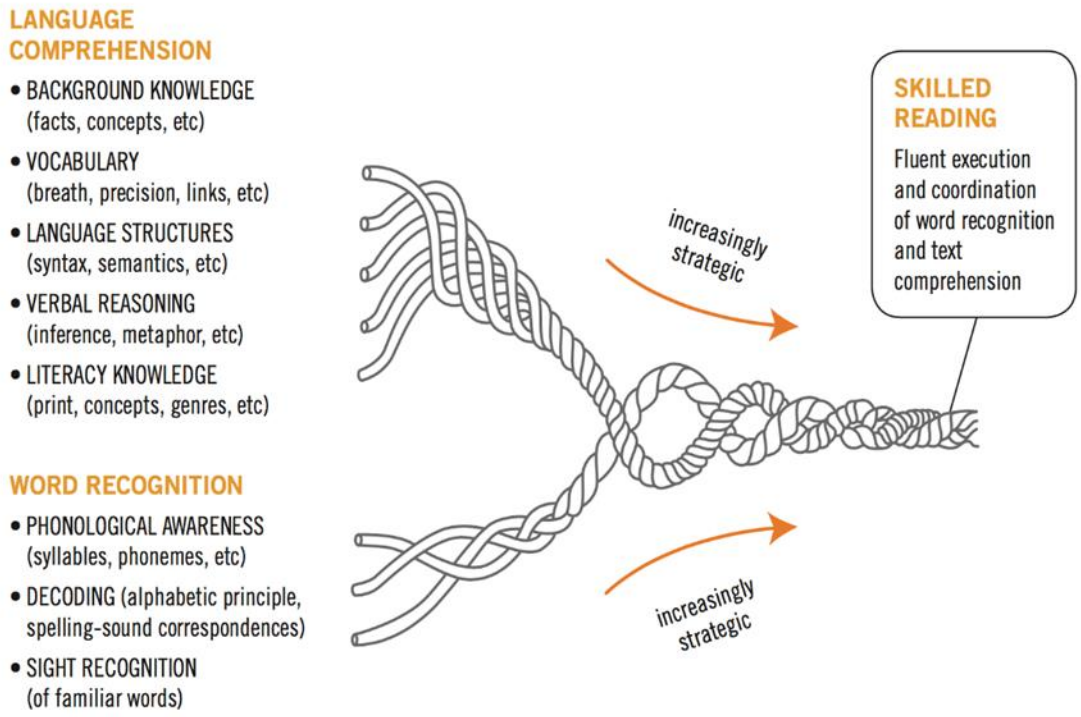
[24] The subsequent section draws heavily on Funda Wande project documentation.
[25] For example, compare the "e'' sounds in "bell", "be", and "the". Similarly, the phoneme /f/ is written as "f", "ph" and "gh" and the words "first", "graph" and "laugh".
[26] Similar language differences also exist within African languages, for example between the Nguni language family (like isiXhosa and isiZulu) with its conjunctive orthographies (i.e. longer words, and fewer words per sentence), and the Sotho language family (like Setswana) with its disjunctive orthography (i.e. more, shorter words per sentence) (Spaull et al, 2020).
[27] See discussion in Spaull and Pretorius (2019: 150-151).

Reading fluency: the speed, accuracy and "naturalness" of intonation of learners reading aloud. Fast and accurate readers, in turn, can concentrate on comprehending what they are reading, transitioning from "learning to read" to "reading to learn" (Funda Wande, 2018).

However, Snow and Kim (2007) note that decoding skills can only play a limited role within the larger task of learners acquiring a broader vocabulary and eventually reading with comprehension. In conjunction with decoding, learners must have a repository of word knowledge (or vocabulary), on which they can draw in order to read with meaning. Intuitively, one can think of reading as the product of learners decoding ability and their vocabulary (Taylor et al., 2019). Vocabulary, in turn, is to a great extent determined by learners' greater socioeconomic context and the quality of language exposure that they have outside of the classroom (Spaull and Pretorius, 2019).

The second main theme of Funda Wande training content focuses on the importance of **reading comprehension**. It emphasises the "significance of reading comprehension skills in enabling individuals to lead rich academic, professional, and personal lives" (Funda Wande, 2018: 22). The framework used is that of Scarborough's Reading Rope in figure 1 below (Scarborough, 2001). This depicts a process whereby *reading* comprehension is a simultaneous interaction of two sub-sets of skills that learners need to have mastered: i) decoding (in the broader sense, referred to as word recognition in figure 1) and ii) a general understanding of the language itself (language comprehension).

FIGURE 1: GETTING TO READING COMPREHENSION - SCARBOROUGH'S READING ROPE



The final component focusses on learner's **affective response** to reading. The theory is simple: children who enjoy reading are more likely to engage in reading for pleasure and thus become better readers (Guthrie et al. 2007). The programme places emphasis on the oft neglected role of learners'

affective (or emotional) response to reading material and how it links to their motivation to read - placing this at the heart of the intervention's content foci.

Furthermore, the provision of whole class sets of quality learning materials ensures that learners have access to high quality reading materials and that their classrooms are immersed in print. The supply of high quality, contextually relevant reading inputs are motivated by the intervention goal of "catching" learners early in their developmental paths by getting them to read for enjoyment. The intervention emphasises the importance of reading role models and teachers' knowledge of the relevant literature available to their learners, as well as identifying and building self-efficacy among learners who struggle with reading. Particularly, Funda Wande course materials require teachers to "be aware of the fiction, poetry, and nonfiction that offer young children opportunities to become familiar with print, to gain information and new perspectives, and to be entertained" (Funda Wande, 2018: 27).

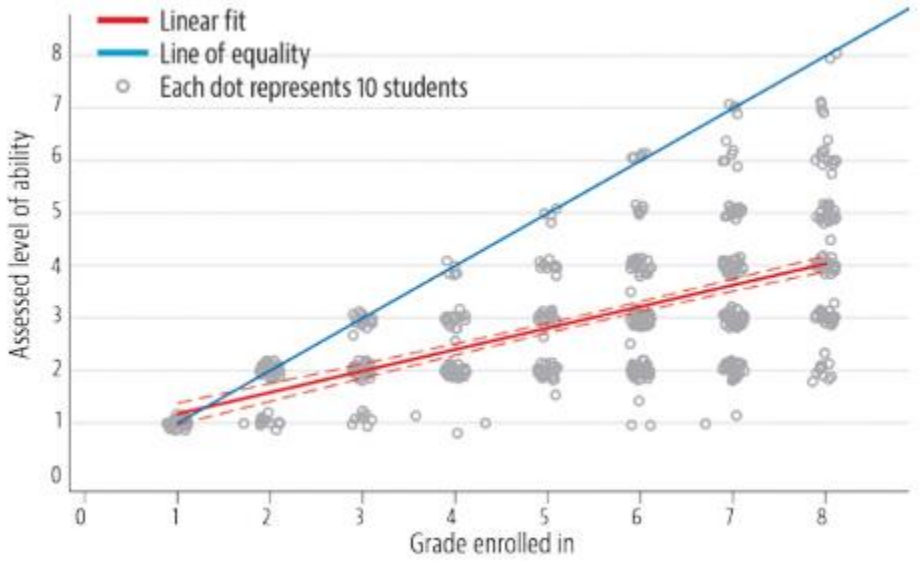
3.2.2.3 TARGETING TEACHING TO STUDENTS' ACTUAL LEARNING LEVELS

Learners who fall behind often learn very little in school if the level of classroom instruction is considerably above their learning level (Banerjee and Duflo 2012, Pritchett and Beatty 2015) [28]. Illustratively, from the Indian case (Muralidharan and Singh, 2019[29]), figure 2 demonstrates a common pattern in developing country education systems: a mismatch between actual learning levels and the level of the curriculum prescribed and actual instruction (Muralidharan et al., 2019). A large proportion of learners fall behind the curricular-prescribed level of learning early in their schooling careers yet are mechanically promoted to further grades. In turn, these learners fall further behind the level of material covered in their textbooks. This results in considerable heterogeneity in learning levels of learners *within the same class*, with the within-class disparities worsening as lagging learners move up the grade levels.

[28] See Banerjee et al. (2017) for an overview of the "Teaching at the Right Level" – type interventions that have developed largely in response to this general observation, with these programmes now effectively implemented at scale across multiple countries in the developing world.

[29] A similar graph for South African learner numeracy levels, but disaggregated by school wealth quintile, is available from Spaull and Kotze (2015:21).

FIGURE 2: THE INCREASED MISMATCH BETWEEN CURRICULUM PRESCRIBED CONTENT AND LEARNERS ACTUAL LEARNING LEVELS



Note: Figure 2 is based on baseline data from more than 5,000 learners in Rajastan, India, in 2017 (data subsequently used in Muralidharan et al., 2019). It shows the estimated level of learner achievement (determined by the Mindspark Computer Asssisted Learning programme) plotted against the grade they are enrolled in.

For Funda Wande learners, the **reading and print materials** component of the intervention provide the opportunity to practise and consolidate new knowledge every step of the way. The Vula Bula graded readers were developed originally and specifically to be relevant in South African African languages (i.e. not adapted and/or translated from English original versions). These readers have formed part of the foundational inputs of other recent structured pedagogic interventions and educational interventions more broadly in South Africa, providing a text that is targeted at the right level and progresses in line with the natural phonic progression of the respective African languages. Similarly, the additional graded reading aides (posters and phonics flashcards) provided are similarly levelled to learners' levels of reading progression at different points in their developmental trajectories (Funda Wande, 2018).

The Funda Wande training materials place a great deal of emphasis on the importance of early remedial intervention and how to go about it in practice. If some children find reading effortful and frustrating, they will most likely not perceive it as meaningful or pleasurable and may therefore become less inclined to actively engage in it. Research into the acquisition of literacy shows that if weaknesses in these areas are overlooked and not remediated, reading problems often persist throughout learners schooling career (Spear-Swerling, 2006). For teachers, the Vula Bula graded readers, handwriting-, baseline assessment- and group guided reading booklets provide a means to continuously assess learner progress and identify when and where learners fall behind. In turn, this should enable teachers to remediate any learning deficits early and effectively.

The provision of group guided reading booklets is intended to facilitate the implementation of the task, making the challenging activity easier to implement and to provide a resource for teachers to track individual learners' progression. The EGRS I study highlighted the important role played by teachers adopting curriculum prescribed teaching techniques that are oft-neglected because of their

more technically demanding nature, like group-guided reading (Cilliers et al., 2019). The activity rarely occurs in classrooms in status quo schooling environments.

The idea is for teachers to group learners according to their reading proficiency levels (in small groups of six to ten learners) and differentially provide reading passages of different difficulty levels to groups based on their respective reading proficiency levels. In turn, this provides learners with both more individualised attention and better targeted instruction. However, the activity requires a high level of pedagogical competence on the part of the teacher: being attuned to the reading proficiency levels of all the learners in their class, having the requisite knowledge of the early grade reading literature to match a specific text to the reading level of a group of learners, and having the classroom management skills to implement this in large classrooms.

3.2.3 SHIFTING TEACHER INSTRUCTIONAL PRACTICE

3.2.3.1 TEACHER COACHING AND LEARNING-BY-DOING

Teacher and HOD training materials systematically cover the disciplinary-, pedagogical and curriculum knowledge contained within the Funda Wande course materials. A common element of teacher professional development programmes is a focus on imparting knowledge. Nevertheless, teaching itself is a skill, something that must be developed through iterative practice and learning-by-doing (Kennedy, 2016; City et al., 2009). Therefore, the coaching component of the programme helps teachers translate the knowledge components of the intervention into in-class pedagogical competence[30].

New methods are often challenging to adopt and require the breaking of old habits - whilst simultaneously learning and concretising new ways of doing. This challenge can be amplified if teachers are provided with little to no guidance, support, or feedback. Some measure of professional accountability and support facilitate transition to a new teaching practice. **One-on-one instructional coaches** fulfils the role.

Concurrently, the recurring professional interaction should foster new expectations and personal development goals on the part of the teacher – fostering a sense of professional accountability to accomplish these goals. In terms of the support provided by the coach, there is also reason to believe that part of the effectiveness of the teacher-coach interaction relies on how the affective state of the teacher is impacted (Hargreaves, 2005; Fullan, 2007). Teacher motivation, self-efficacy and the extent to which they cope with emotions of frustration, anxiety, insecurity, etc. provide further channels through which the coach-teacher relationship can impact on instructional change.

In addition to classroom visits, the coaches also have **afternoon phase workshops** (once per school visit) and occasional larger **cluster meetings** with teachers from the same community. These sessions entail activities like working through the lesson plans and focussing on strategies to deal with common challenges faced by the teachers. This also provides teachers with the opportunity to learn from the successful practices and failures from those in their cohort, effectively creating localised communities of professional practice (Chauraya and Brodie, 2017).

[30] A familiar argument in the educational literature is that teachers are more likely to adopt new instructional approaches based on shifts in their own attitudes and beliefs, which are in turn based on first hand experiences of shifting learner outcomes (Guskey, 2002).

3.2.3.2 *Necessary Support factors for effective teacher coaching*

The success of the instructional coaching component of the intervention therefore relies on the capacity of coaches themselves and the nature of their relationships with the respective teachers. From a practical programme perspective, expectations on the role of coaches are clearly outlined to ensure that it functions in precisely the role intended by the programme (Fullan and Knight, 2011). Coaches also play an important role in managing the relationships with the schools, observing certain protocols to ensure good relationships with school principals and teachers alike.

Given that the success of the intervention relies on the levels of buy-in and participation from teachers, this also requires that the programme carries legitimacy, purposeful time allocation in teaching schedules and thus the perceived support of authorities (school principals, district-level and provincial-level officials). Fleisch et al (2017b) argue that programme compliance will be low in the absence of governmental leadership and buy-in; if teachers do not perceive the programme as being a clear priority and part of their core teaching responsibilities. The importance of teacher buy-in has also been highlighted elsewhere, for e.g. India and the USA[31]. Banerjee et al (2017) document the importance of state-bureaucratic and researcher-implementer relationships in effectively scaling Teaching at the Right Level (TaRL) programmes in India., Even when teachers believed in the usefulness of the programme and the materials it provided, they largely did not adopt methods when these practices were not viewed as part of their core responsibilities (Banerjee et al, 2017). Only once the organisational buy-in was achieved, through emphasising that the state-government was supportive of- and responsible for the programme's implementation, as well as the programme receiving its own government-mandated time-slot for implementation in school hours, did the programme lead to teacher instructional change.

Since Funda Wande as the programme implementers are a non-governmental organisation (NGO), various measures were taken to ensure the support and buy-in of the relevant stakeholders[32]. The HOD bursary for the certified course provides further legitimacy.

3.2.3.3 *Integrated support materials to facilitate shifts in teacher practice*

Whereas most teachers might have historically lacked the basic classroom resources, teachers "are now overwhelmed when receiving large volumes of texts in the form of workbooks, big books, posters and graded readers" (Funda Wande, 2018:30). The instructional coaching and structured lesson plans specifically target teachers' ability to integrate these resources into their classroom practices.

Lesson plans are intended to reduce the daily cost of adopting new teaching practices (especially when coaches are not around). The counterfactual scenario for teachers in status quo schooling environments is to conduct curriculum planning amongst themselves and subsequently craft their own lesson plans. Lesson plans were thus designed to reduce teachers' planning and administrative workloads, allowing them to focus on actual teaching and implementing the pedagogical skills that they have been equipped with. The goal is to reduce the mental bandwidth that teachers expend on

[31]Similarly,in a review of 90 evaluations of US educational interventions, the research arm of the US Department of Education found that a lack of implementation was an important contributing factor to why 79 of 90 these interventions did not have positive effects (Coalition for Evidence-Based Policy, 2013).

[32] For example, the ECDoE and district managers were involved from the project conceptualisation stage, helping to identify the schools that would be eligible to apply to become part of the intervention.

lesson planning, to serve as a daily reminder and prompt of newly learnt techniques and to provide a framework within which a new routine and habits can form. On-line resources and pre-loaded flash drives with the full set of Funda Wande videos and multimedia resources further improve the ease of access to materials that teachers can use to learn and/or refresh new pedagogical skills.

Lesson plans are aligned to the South African national curriculum (CAPS). Therefore, they provide teachers with a guide to effective lesson pacing, sequencing and a structured approach towards current curriculum coverage. Given that most teachers in South African classrooms cannot keep up the pacing and coverage as prescribed by curriculum benchmarks (Hoadley, 2012; Taylor and Taylor, 2013), the blueprint provided offers another avenue along which teachers can more productively allocate time across teaching activities.

The lesson plans in the Funda Wande intervention are fairly prescriptive, laying out step-by-step what is expected from teachers in each respective lesson. Lessons plans explicitly require that teachers "(t)ry to stick to the lesson plans as much as possible" (see course materials in appendix figure B1 to B11). Despite ongoing debates around optimal levels of teacher autonomy[33], findings from Cilliers et al. (2019) provide evidence that fairly prescriptive lesson plans can indeed form part of effective intervention in low resource-, low capacity South African classrooms[34].

4 EVALUATION DESIGN

4.1 RESEARCH 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, the aim is to investigate the impact of the intervention on both foundational reading skills and reading comprehension in the learners' home language.

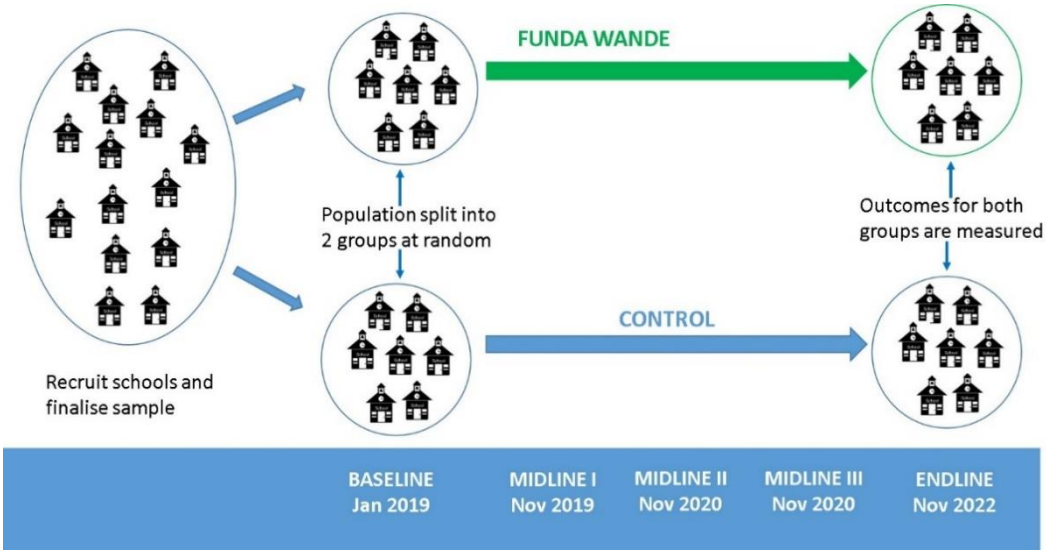
4.2 EVALUATION METHODOLOGY OVERVIEW

Baseline reading assessments were conducted with learners in all schools before the programme started (in January 2019). The results here are based on the midline assessment completed at the end of the first academic year (November/December 2019) – the first round of follow up of the four planned over the period. The midline results form part of an impact evaluation using a randomised control trial (RCT). Schools are 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.

[33]See Dresser (2012) and Piperet al. (2018).

[34] In the poorly performing no fee public schools in the North West province where the EGRS I study was conducted, teachers reported a particularly shifting both teacher's instructional practice and learning outcomes, with no detectable negative impact on any segment of the student population (Cilliers et al., 2019a).

FIGURE 3: FUNDA WANDE IMPACT EVALUATION DESIGN



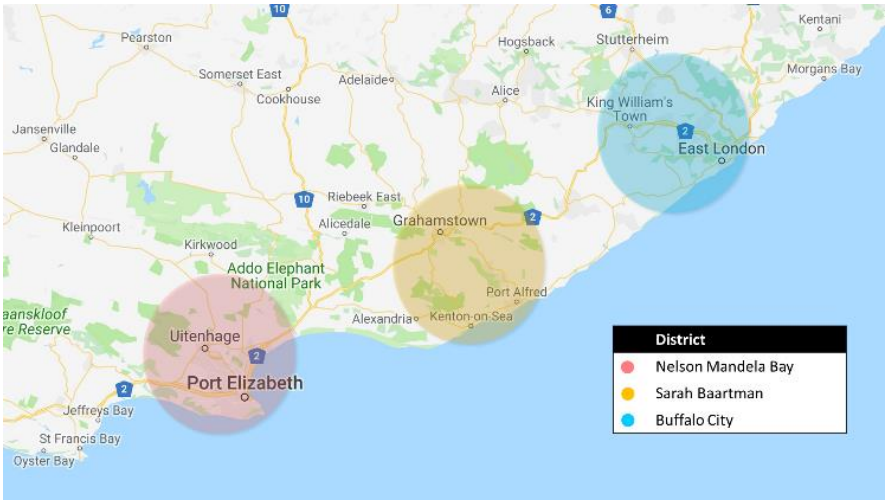
4.3 RECRUITMENT AND RANDOM ASSIGNMENT

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[35].

Funda Wande's approach is to "work with the willing"– to work with schools who 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. Annex III provides further details on the school selection process and how representative the evaluation sample is of the underlying population of schools in the Eastern Cape.

[35] These criteria were decided in collaboration with the ECDoE.

FIGURE 4. LOCATION OF FUNDA WANDE IMPACT EVALUATION SCHOOLS



The final baseline sample is comprised of 29 control schools and 30 treatment schools, with schools randomly assigned to treatment or control within each of the three strata (approximately the three educational districts – see Annex III). Within each school, one grade 1 class and one grade 2 class were randomly selected; and within each of these selected classes, 10 learners were randomly selected.

4.4 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 Funda Wande technical experts and the evaluation's Principal investigator[36].

Appendix table C1 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, one would expect floor effects (i.e. many learners scoring zero) in many of the core EGRA subtasks, particularly for grade 1 learners. A range of subtasks are thus employed, including pre-literacy measures, in order to ensure that there is good discrimination between learners at baseline and at midline. Annex IV in the appendix further outlines the rationale for how, why and when certain tasks were assessed.

[36] This instrument development process drew heavily upon 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) and NORC (Vocabulary).

At the end of the learner assessment, a short interview was conducted 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. All assessments and interviews were conducted entirely in isiXhosa by isiXhosa-home-language enumerators.

4.5 BASELINE SAMPLE DESCRIPTION

Overall, 1187 learners were assessed at baseline[37]. Tables 1 to 3 provide a description of the sample at baseline. The sample is roughly evenly split between boys and girls. The average age of Grade 1 learners is 6 years and 5 months, whilst Grade 2 learners are on average 7 years and 6 months old. Almost ten percent of the Grade 1 learners are repeating their grade, with less than five percent of grade 2 learners repeating.

TABLE 1: LEARNER SEX, AGE AND GRADE REPETITION BY GRADE

	Grade 1	**Grade 2**	**Total**
Girl	49%	51%	50%
Age	6 years 5 months	7 years 6 months	6 years 11 months
Repeating grade	9.7%	4.6%	7.1%
Observations	595	592	1187

Table 1 displays some general learner- and home level characteristics. As a proxy for learner health and nutritional wellbeing, the height-for-age z-score is 0.4 standard deviations below the mean for the healthy reference population. At the extreme end, approximately five percent of the learners in the sample are classified as stunted (have a z-score below negative two). The majority of learners live with both their parent (61 percent), whilst children from single parent households are most likely to be living with their mother. Limited information on household wealth was also gathered – with learners reporting that their households have on average just over six of the nine basic assets and services listed. Only a third of learners have access to a computer at home, and only half come from households that own a vehicle.

[37] 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, the research team 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 2: LEARNER AND HOME CHARACTERISTICS

	Mean	Std Dev.
Height-for-age z-score	-0.38	1.01
Stunted	0.05	0.21
Lives with both parents	0.61	0.49
Lives with mother	0.90	0.30
Lives with father	0.63	0.48
No books in home	0.65	0.48
Radio	0.71	0.45
Mobile	0.99	0.10
Electricity	0.97	0.18
Television	0.94	0.23
Computer	0.32	0.47
Fridge	0.93	0.26
Toilet	0.60	0.49
Bicycle	0.33	0.47
Vehicle	0.51	0.50
Count of household assets and electricity access	6.28	1.63

A comprehensive discussion on learners' access to print resources, both at home and in the classroom, is provided in the baseline report (Ardington, 2019). Suffice to summarise here that the Funda Wande evaluation schools follow the wider trend in low resource schools of a lack of access to reading materials and limited opportunities to engage with text in a meaningful way (both at school and at home). More than half of learners do not have access to a library at school (either on site or mobile), very few learners have access to "print rich" classrooms, the number of readers available per class are generally insufficient for the number of learners in the class. Table 2 above suggests a similar lack of access to reading material at home, as almost two thirds of learners (65 percent) report that they have no books other than schoolbooks to read at home.

Table 3 provides a sense of learners' general lack of pre- and early literacy skills at the start of their respective grades – summarising their scores on select grade relevant EGRA tasks. At the start of Grade 1, learners could identify less than six correct letter sounds per minute on average, whilst half of the learners could not identify a single letter sound correctly. Almost no Grade 1 learners could identify more complex digraphs and trigraphs. The phonemic awareness subtask requires learners to identify

the starting sound of words and progresses to more challenging word segmentation tasks. Almost half (49%) of the Grade 1's scored zero, implying that these learners could not identify the start sound of simple consonant-vowel-consonant-vowel (CVCV) words.

By Grade 2, learners could at least complete fundamental decoding tasks with some level of success. They scored significantly better on the two letter sound tasks on average (correctly identifying 29 letter sounds and almost 9 digraphs and tri-graphs correctly per minute), but more than half of these learners could still not identify a single digraph or trigraph correctly. Grade 2 learners had an average phonemic awareness score of four out of ten at the start of the year. For oral reading fluency, learners were able to read just over seven words correctly per minute from paragraph of connected text (short story) – whilst 44 percent could not read a single word from the passage. Grade 2 learners answered on average only four (out of 15) comprehension questions correctly, with almost half unable to answer a single comprehension question correctly

TABLE 3: SUMMARY OF SELECTED BASELINE EGRA SCORES, BY GRADE

	Grade 1			Grade 2		
	Mean	**Std Dev**	**% zero**	**Mean**	**Std Dev**	**% zero**
Letter sounds p/m	5.5	9.3	51%	28.5	19.9	12%
Di-/ Trigraphs p/m	0.2	1.3	96%	8.5	11.3	51%
Phonemic Awareness	1.5	1.8	49%	4.3	2.3	9%
Oral Reading Fluency				7.4	9.4	44%
Reading Comprehension				4.2	4.7	46%

Figure 4 shows the distribution of letter sounds scores for Grade 1's, indicating the percentage of Grade 1's who could correctly identify the number of letter sounds per minute corresponding to the respective bins (by treatment status). Seventy percent of Grade 1 learners could identify no more than five correct letter sounds per minute at baseline. At the other end of proficiency, six learners (one percent) could identify 40 or more letter sounds per minute correctly.

FIGURE 4: HISTOGRAMS OF GRADE 1 BASELINE LETTER SOUND RECOGNITION, BY TREATMENT STATUS

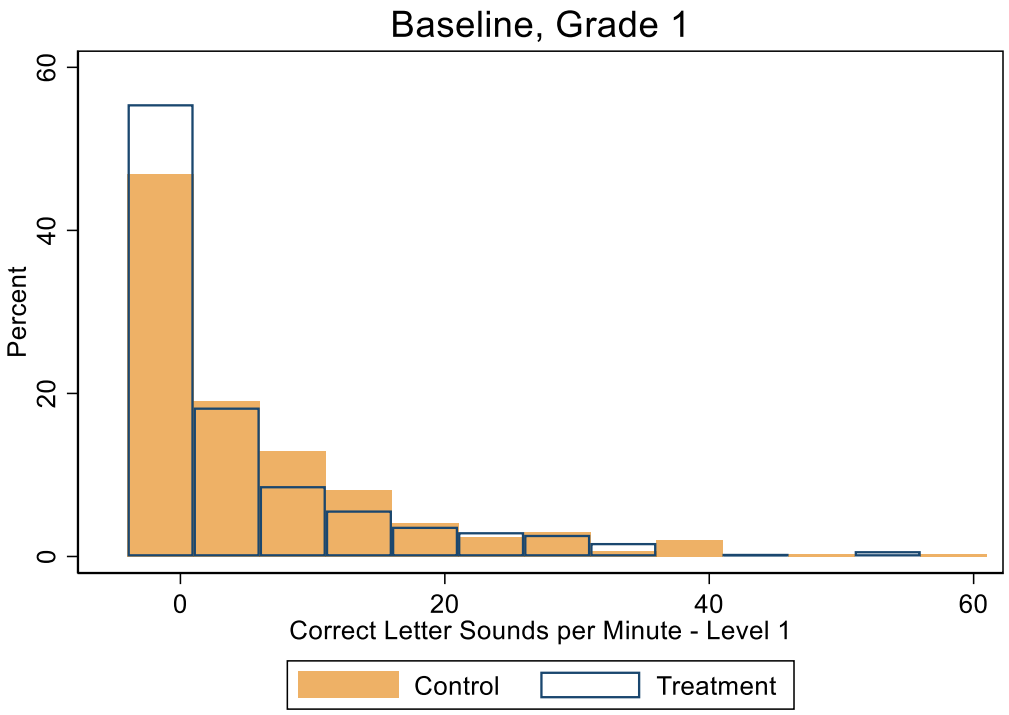
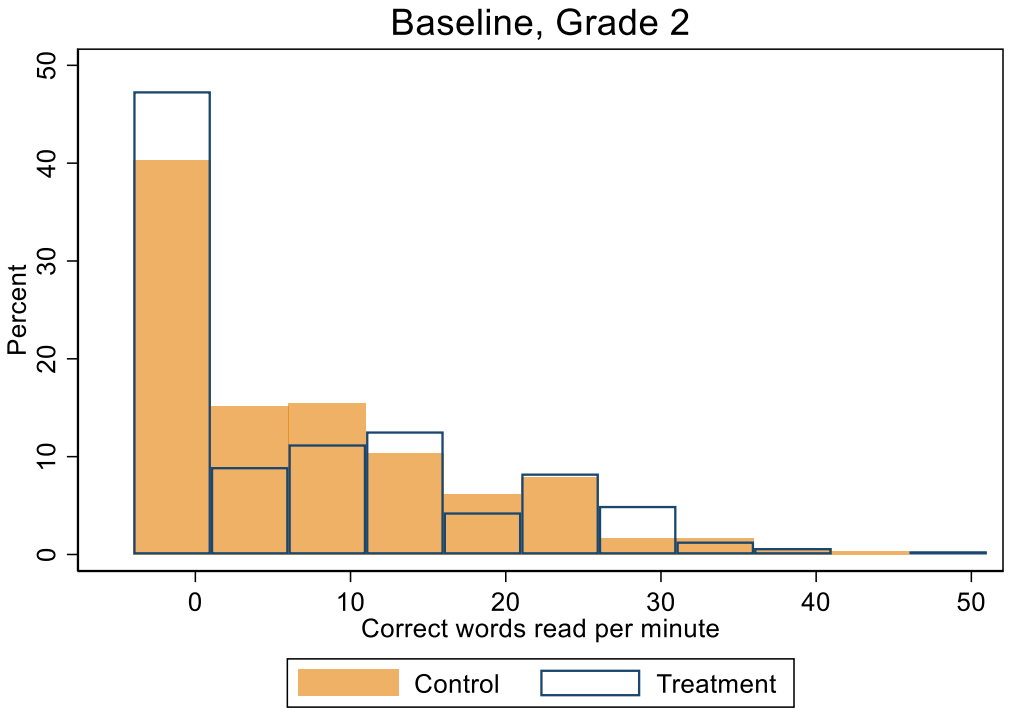


Figure 5 indicates a similar distribution for Grade two learners' oral reading fluency. More than half (56 percent) of Grade 2 learners could read no more than ten words correctly per minute from the short passage, whilst three percent of learners could read more than 30 words correctly per minute. For both task-grade pairs, it is clear that the average scores for the group are inflated by a small share of learners who are fairly proficient in the sub-task task, whilst a majority of the group scores low-single digits or zero. As one would expect, figures 4 and 5 also show that control- and treatment school learners EGRA scores follow similar distributions.

FIGURE 5: HISTOGRAMS OF GRADE 2 BASELINE ORAL READING FLUENCY, BY TREATMENT STATUS



Random assignment of schools to treatment and control eliminates any possible selection bias (Angrist and Pischke, 2009, Athey and Imbens, 2016). It ensures that any subsequent differences in learners' midline reading proficiency by treatment status can be attributed to the Funda Wande intervention (i.e. the treatment effect). Done correctly[38], random assignment ensures that schools are balanced on observable and unobservable characteristics in expectation. Any incidental differences between the treatment and control groups at baseline therefore occur by chance. For this reason, it is not necessary to conduct an array of individual tests to assess whether baseline imbalances are statistically significant (see Athey and Imbens, 2016; Bruhn and McKenzie, 2009). Nonetheless, given that sample balance checks are common practice in development economics, appendix tables A1 to A3 report balance tests for the full array of observable learner characteristics at baseline. As one would expect, these balance tables demonstrate the similarity between the treatment and control groups in terms of learner and household characteristics and reading skills -with only one of the 47 significance tests (2%) indicating a statistically significant difference at the 5% significance level (no more than expected to occur by chance) [39].

4.6 MIDLINE SAMPLE, ATTRITION AND BALANCE

Overall, 94 percent of learners assessed at baseline were re-assessed at midline[40] (Table 4). The predominant reason for learner attrition were either that learners were absent on the days that

[38] In this case, using the statistical programme Stata.
[39] See baseline report (Ardington, 2019) for a more extensive discussion on baseline characteristics.
[40] 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. Both the Baseline report and the Midline Field Report are available upon request.

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[41]. In 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 that prevented assessments from taking place, or the assessment was not captured due to technical challenges in data collection[42].

TABLE 4: MIDLINE ASSESSMENT 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 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.

Attrition has two potential impacts on the RCT. The first is a small reduction in statistical power with a slightly smaller sample. However, the power calculations behind the sample design are based on fairly conservative assumptions, implying that the small reduction in sample size here is not of great concern[43]. The second is the potential for selection bias to be introduced into the sample, thereby

[41] This accounts for a further two percent of the baseline sample. Researchers 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, there was no choice but to remove this school.

[42] Table 4 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.

[43] See the baseline report for the original power calculations (Ardington, 2019). These power calculations were particularly conservative given i) the high levels of correlation between baseline and midline measurements of learners reading proficiency (especially for Grade 2 learners) and ii) the relative homogenous nature of schools in the sample (implying that most of the variation in outcomes is between learners within schools, and not between schools themselves). Viewed in conjunction with the fact that there is very little attrition at this stage of the evaluation, this implies that treatment effects can be fairly precisely estimated.

threatening a key strength of the RCT methodology: the internal validity of the estimated programme impacts.

The randomisation process faced no challenges and the midline analysis sample remained balanced – implying that impacts can be reliably estimated. Column (1) in Table 5 shows the regression of treatment status on whether or not a learner attrited, taking into account the experimental design[44] (Athey and Imbens, 2016). The overall attrition rate for the learner sample was six percent[45]. Learner attrition was slightly higher amongst learners in treatment schools as compared to learners in controls schools (seven percent versus five percent – see table 5 below [46]). There is no statistically significant differential attrition between learners in the two groups. As is to be expected from previous evaluations (e.g. Cilliers et al., 2019), teacher attrition was slightly higher at eight percent[47]. Although the attrition rate was higher for teachers in treatment schools than control schools (10 percent versus seven percent), this difference was also not statistically significant.

TABLE 5: TESTING FOR DIFFERENTIAL ATTRITION

	Learner Attrite	Teacher Attrite
Treatment	0.030	0.035
	(0.032)	(0.053)
Control Attrition	0.046	0.069
Observations	1,187	1,187
R-squared	0.054	0.162
Strata FE	YES	YES

Note: Standard errors reported in brackets.

To confirm that the midline sample is still balanced, table 6 reports the standardised mean differences in baseline outcomes and characteristics for the midline sample (Athey and Imbens, 2016; Imbens and Rubin, 2015: Chapter 14). Rather than reporting statistical significance, the focus is on the size of the differences between the groups (and whether any imbalances affect the outcomes of interest)[48].

[44] In other words, taking into account the sub-district (or strata) within which schools were randomly assigned and clustering standard errors at the school level.

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

[46] From table 5, column (1): there is a three percent mean difference in attrition rates between learners in the two groups. The low overall attrition and minimal differential attrition comfortably fall within what is considered acceptable to reliably estimate the programme's impacts. For example, following common education evaluation guidelines from What Works Clearinghouse (WWC) (2020:10), the low attrition rates comfortably fall within the conservative limits for low expected bias meet the highest possible WWC standards.

[47] 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 percent.

[48] The variables reported in table 6 are all included as covariates in the subsequent analysis because i) they are predictive of midline reading proficiency (improving precision), ii) they were incidentally imbalanced at baseline, and/or iii) they display slight imbalances for the midline analytical sample (based on the effect sizes).

Standardised mean differences (of effect sizes[49]) provide a scale invariant measure to assess whether the two groups are still equivalent at midline. As expected, the differences between treatment and control learners on all variables are within the limits to satisfy equivalence (Imbens and Rubin, 2015: Chapter 14, WWC, 2020)[50]. Variables for which there are slight imbalances are included as covariates in the subsequent analyses. Results displayed are for the midline sample of learners, but the same results hold for the full baseline sample (equivalent up to the second digit after the decimal point – not displayed). This is to be expected, as the sample was balanced at baseline (see appendix tables A1 – A3)[51] and as there is little learner attrition between assessments.

TABLE 6. CHECKING MIDLINE EQUIVALENCE – DIFFERENCES IN MIDLINE SAMPLE OF LEARNERS' BASELINE TEST SCORES, -CHARACTERISTICS AND -HOME ASSETS (BOTH GRADES)

	Treatment		Control		p-value	Pooled	Effect
	Mean	s.d.	Mean	s.d.		s.d.	size
Common tasks							
Letter Sounds per minute	17,31	19,78	17,49	19,08	0,91	19,42	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	0,86	1,50	0,01
Receptive Listening Comprehension	9,63	0,74	9,50	1,01	0,13	0,88	0,15
Phonemic Awareness	2,87	2,51	2,98	2,52	0,62	2,51	0,04
Expressive Vocabulary	11,28	3,80	11,69	4,04	0,31	3,93	0,11
Write letters	2,70	1,93	2,89	1,80	0,25	1,86	0,10
Grade 1 only tasks							
Word Choice	1,73	1,74	1,90	1,60	0,36	1,67	0,10
Rapid Automatized Naming	36,99	12,00	37,63	11,59	0,59	11,79	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	4,33	1,36	0,49	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	4,20	4,83	4,34	4,59	0,83	4,71	0,03
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							

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

[50] 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 are included as controls in the model estimating programme effects to satisfy equivalence Variables for which such adjustments are required include 11 of the 18 sub-tasks, learner age, whether learners have non-academic books at home, whether their household owns a computer or some form of motor vehicle, or has a toilet inside their home. Variables with effect sizes in this range are displayed in red. These variables are 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 not considered to be equivalent any more (WWC, 2020).

[51] None of the 47 baseline outcomes and covariates are statistically different at the p=0.05 level.

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	0,37	0,50	0,04
Age in months	84,16	10,63	83,36	10,21	0,32	10,42	0,08
Height for age z-score	-0,37	1,04	-0,38	0,98	0,90	1,01	0,01
Household assets							
Books other than schoolbooks to read at home	0,33	0,47	0,38	0,49	0,10	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	0,69	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	0,33	0,47	0,38	0,49	0,10	0,48	0,12

5 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. The two sections below provide a descriptive overview of midline outcomes in both groups before moving on to quantifying the impact of the Funda Wande intervention.

5.1 DESCRIPTIVE STATISTICS

The sections below outline how the respective reading skills developed over time in status quo schooling environments (section 6.1.1.), as well as a descriptive comparison of the eventual reading proficiency of learners in treatment schools relative to the comparison group (section 6.1.2. and 6.1.3)[52].

5.1.1 DEVELOPMENT OF READING SKILLS IN THE CONTROL GROUP

An understanding of teacher support and learning conditions in control schools are required to get a sense of what any intervention impacts translate to in a practice. First, in terms of teacher professional development -and support, principal data indicates that all the schools (treatment and control) reported receiving some form of intervention targeted at improving literacy instruction in the three-year period from 2017 to 2019. The National Education Coalition Trust (NECT) structured reading programme in particular seems to be the status quo support intervention for schools in the Eastern Cape. Eighty percent of control schools (and 84 percent of all schools in the sample) report having received training and/or resources from the NECT[53]. Any Funda Wande intervention effects should thus be interpreted relative to the status quo NECT support, which also includes daily scripted lesson plans and materials for reading lessons.

[52] See baseline report (Ardington, 2019: 14-35) for an extensive discussion on learners reading proficiency and the reliability of the reading assessment instruments at baseline specifically.

[53] Similarly, data from classroom observation studies in 20 classrooms in 10 Eastern Cape school in the SPS sample also shows that teachers were universally following the NECT provided lesson plans and reading materials.

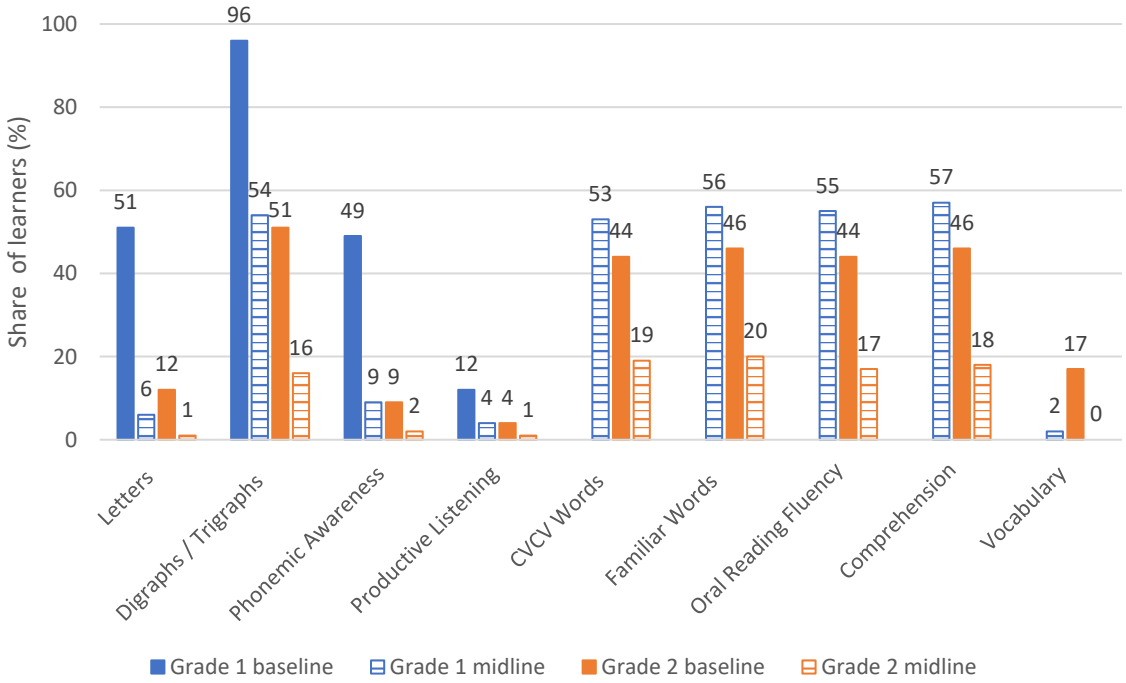
Table 7 presents a summary of the average midline score, standard deviation and percentage of learners scoring zero on each of the EGRA sub-tasks for Grade 1 and Grade 2 learners in control schools. For all the tasks that were conducted at both baseline and midline, the change in the percentage of control group learners scoring zero and the average scores are also summarised in Figures 6 and 7.

TABLE 7: 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		

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.

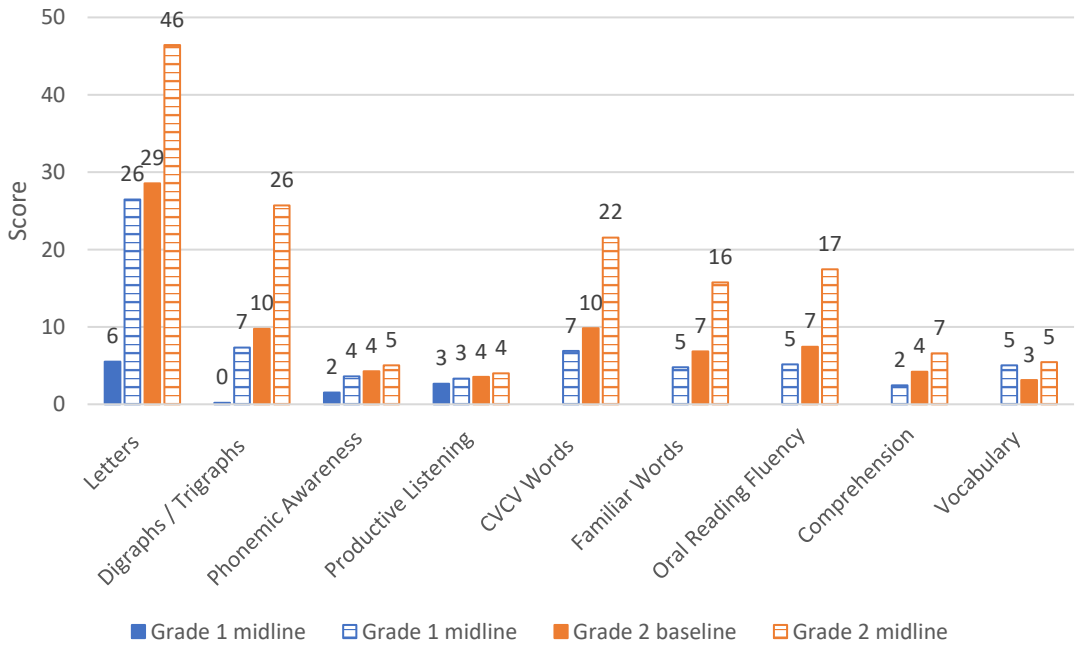
FIGURE 6: PERCENTAGE OF CONTROL GROUP LEARNERS SCORING ZERO AT BASELINE AND MIDLINE



By the end of Grade 1, 50 percent of learners in the control group cannot read a single word (neither four letter consonant-vowel- consonant-vowel (CVCV) words, nor slightly longer familiar words, nor a single word from a paragraph). For those learners who could read some of the passage, the average Oral Reading Fluency (ORF) is 10 words read correctly per minute. Only one in twenty Grade 1's had an ORF of more than 25 correct words per minute.

Almost all Grade 2 learners can correctly identify at least one letter sound and the percentage able to correctly sound at least one digraph has risen from 53 percent at the start of the year, to 85 percent by year end. By the end of Grade 2, the percentage of learners who are unable to correctly read a word from the CVCV words, familiar words, and two reading passages range from 16 to 24 percent. For those learners who could read at least one word of the first passage, the average ORF is 20 words per minute.

FIGURE 7: AVERAGE SCORES FOR CONTROL GROUP LEARNERS AT BASELINE AND MIDLINE



5.1.2 PRIMARY OUTCOME MEASURES

5.1.2.1 COMPOSITE READING PROFICIENCY

In a similar fashion to Cilliers et al. (2019), a composite score of the isiXhosa reading proficiency was constructed based on the different EGRA sub-tasks on which learners were assessed. 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). A baseline composite reading proficiency score is constructed in a similar fashion, but using the EGRA scores for the full baseline sample (i.e. when test scores in the treatment group were still unaffected by the treatment) [54].

[54] 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.

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) (Ardington, 2019).

5.1.2.2 READING COMPREHENSION

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 comprehension sub-tasks in annex II, the additional two minutes of reading time enabled learners to answer a far greater share of the comprehension questions. Table 8 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 as 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.

TABLE 8: HOW MANY COMPREHENSION QUESTIONS LEARNERS CAN ATTEMPT BASED ON THEIR READING SPEED

	Can attempt 1st inferential question (>=13 words read[55])		**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%

The average scores on the comprehension tasks are thus slightly misleading if one is only interested in learner reading comprehension conditional on how fast they read. For example, the average comprehension score for all Grade 2's is 47 percent on the first comprehension task and 41 percent

[55] 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).

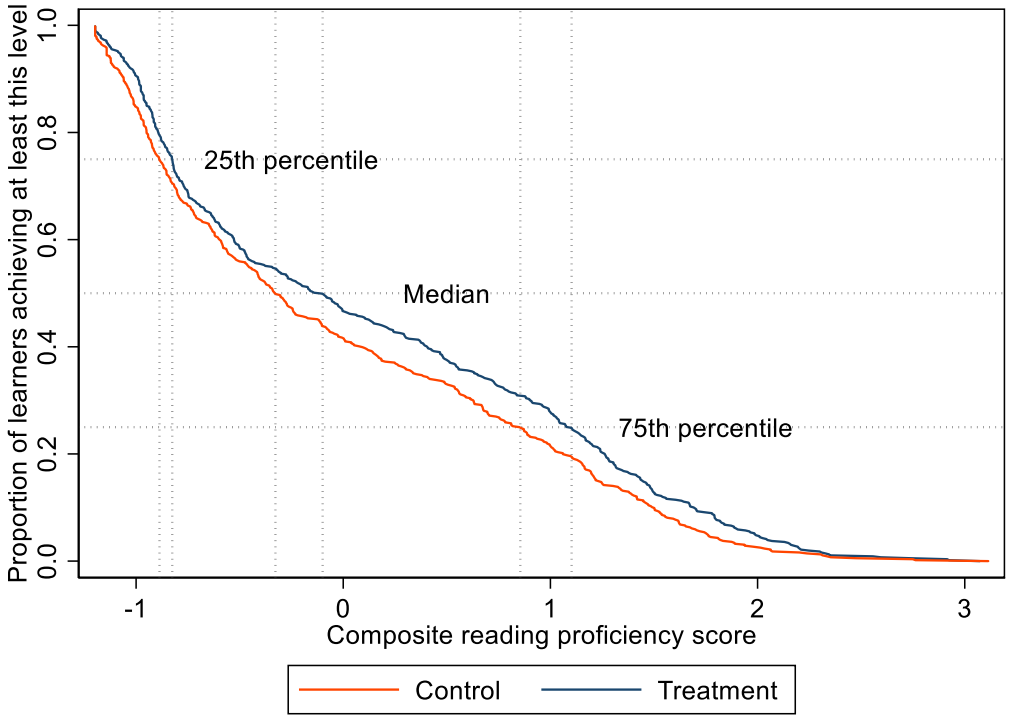
on the second. However, only six in ten Grade 2 learners could attempt all 14 comprehension questions on the first comprehension task, whilst only approximately half (47 percent) of the same group read fast enough to attempt all ten comprehension questions on the second passage. In contrast, for these 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.

Given that challenges in measuring reading comprehension in a context of low levels of reading fluency, an additional untimed sentence choice comprehension task was also included. This task provided only limited discrimination between learners reading comprehension abilities, however. The former two paragraph reading comprehension tasks are therefore the preferred measures of learners reading comprehension abilities. Annex II in the appendix provides further descriptive results and discussion on the three different reading comprehension tasks.

5.1.3 Descriptive analysis of intervention impacts

The distribution of the midline composite score is shown separately by treatment status in Figure 8. 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 standard deviations (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.

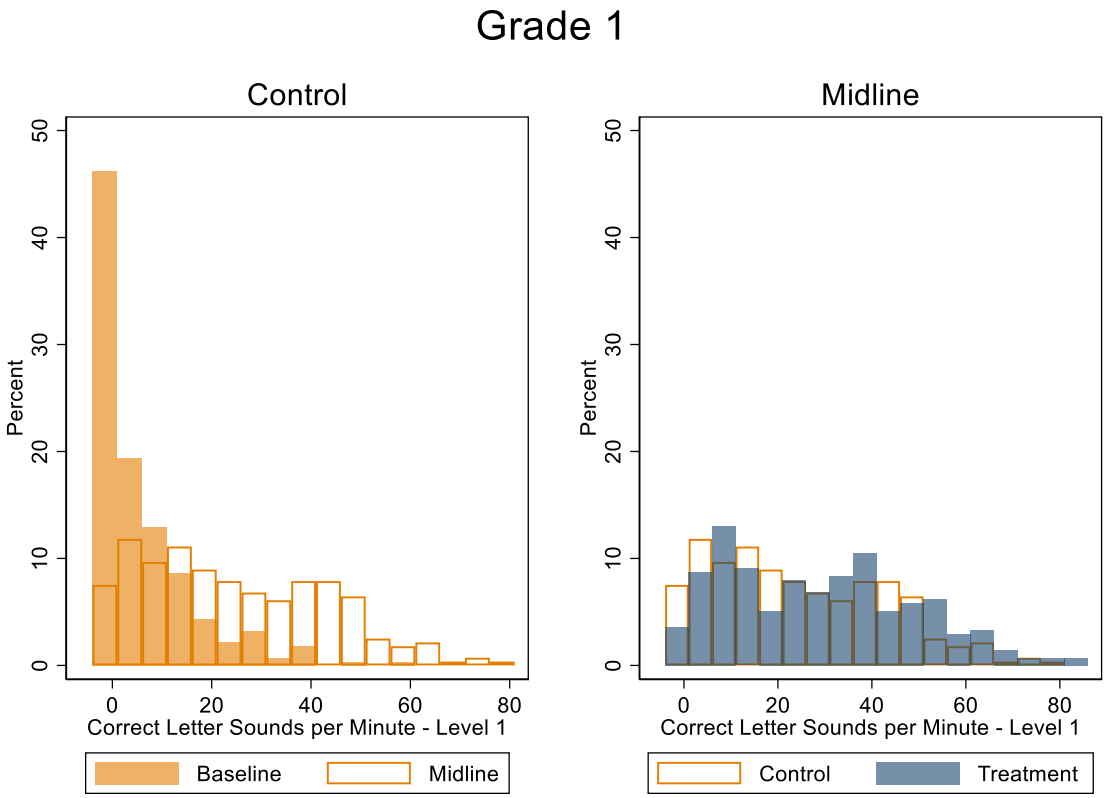
Figure 8: Distribution of midline composite scores by treatment status (full sample of both grades combined)



We next turn to the shifts in the distributions of selected grade relevant tasks. The histograms in figures 9 to 11 below indicate i) how learners' reading abilities in the **control group improved** in the status quo learning environment **over one year (left panels)** and ii) to what extent the **intervention further shifted** the distribution of learner **outcomes by midline**, in addition to what occurred in the comparison group **(right panels)**[56]. In each case the left panel thus 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 9 shows the percentage of Grade 1's who could correctly identify the number of letter sounds per minute corresponding to the respective bins, by treatment status and assessment wave. For example, at baseline 46 percent of Grade 1 control school learners could not read a single letter sound. At midline, the share of control school learners who could not correctly identify a single letter sound decreased to only eight percent. For intervention school learners, however, the share of zero scores was even less at only 4 percent. Overall, figure 9 provides a sense graphically of the extent to which the intervention further shifted the distribution of Grade 1's letter recognition scores to the right, and how significant this is relative to the growth that occurred under status quo learning environments.

FIGURE 9: HISTOGRAMS OF SHIFTS IN GRADE 1 LETTER SOUND RECOGNITION ABILITY

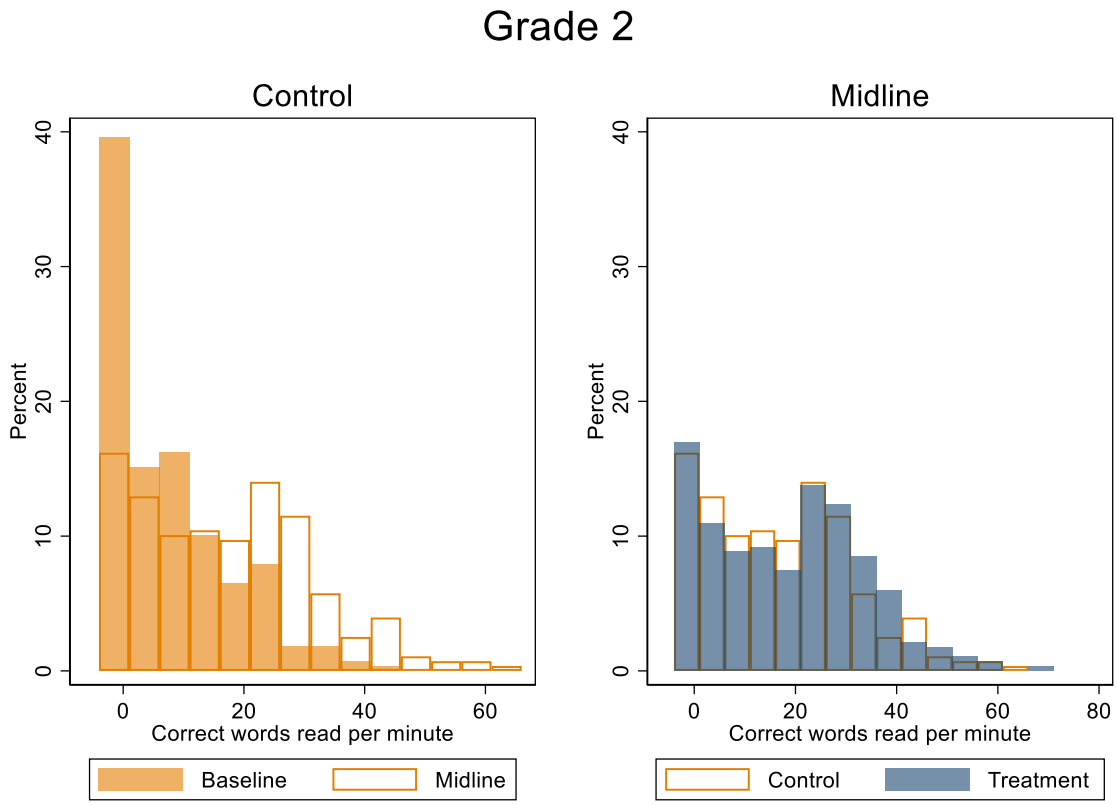


NOTE: BIN-WIDTH = 5 LETTER SOUNDS, WITH A SEPARATE BIN FOR ZEROS ONLY

[56] Summary information on the distribution of each sub-task by grade is displayed in Appendix table A5.

Figure 10 and 11 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 10, 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 words per minute (from 71 percent to 39 percent). 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 (47 percent versus 41 percent in control schools).

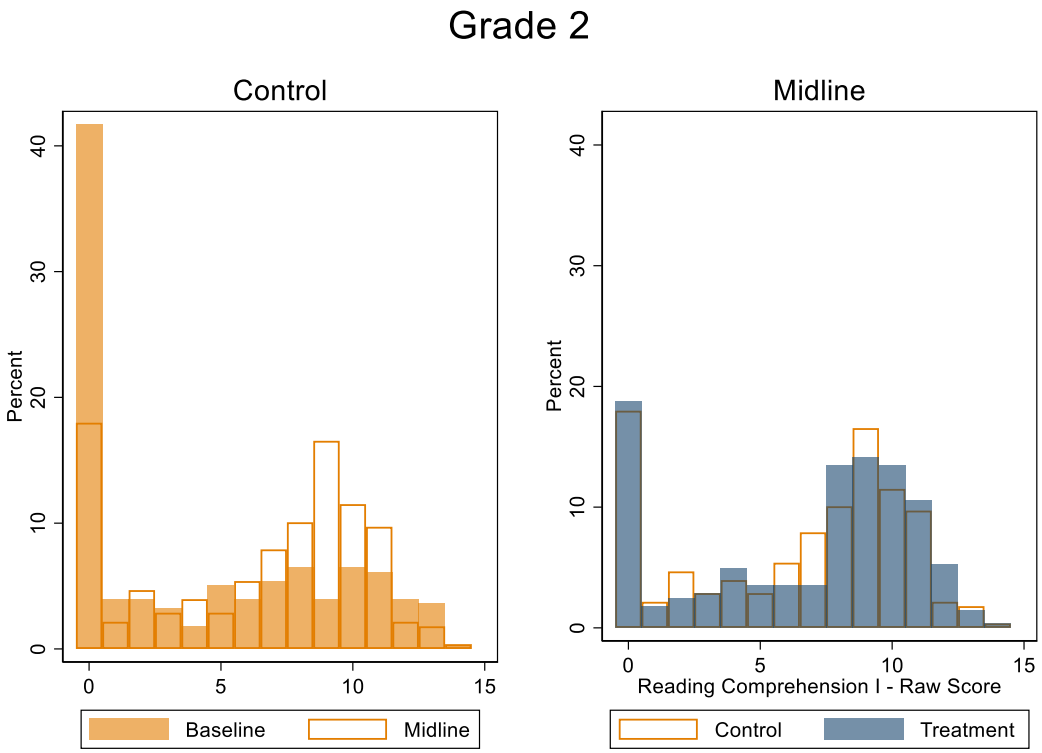
FIGURE 10: HISTOGRAMS OF SHIFTS IN GRADE 2 ORAL READING FLUENCY ABILITY



NOTE: BIN-WIDTH = 5 LETTER WORDS, WITH A SEPARATE BIN FOR ZEROS ONLY

For Grade 2 learners' reading comprehension score, figure 11 indicates that there were significantly less learners in control schools scoring zero on the comprehension task at midline (declining from 42 to 18 percent), in conjunction with a much greater proportion of learners answering at least half of the questions correctly (increasing from 36 to 60 percent over the year). In comparison to the control schools, intervention school learners had a higher share answering more than two thirds of the comprehension questions correctly at midline. Thirty-one percent of intervention school learners scored 10 or more out of 14, compared to 26 percent in the control group. In sum, figures 10 and 11 graphically illustrate two themes in 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. Second, intervention schools have a greater share of learners obtaining relatively high scores on these tasks at midline.

FIGURE 11: HISTOGRAMS OF SHIFTS IN GRADE 2 READING COMPREHENSION ABILITY



NOTE: BIN-WIDTH = 1 COMPREHENSION SCORE

A simple comparison of means on the sub-tasks provides an initial indication of programme impacts. Table 9 provides the differences in means by treatment status, for each sub task and by grade[57]. 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. Focussing on full sample results, 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. The next section investigates whether these differences are statistically significant.

[57] See appendix table A5 for a more detailed information on the distributions of midline scores by treatment status.

TABLE 9: DIFFERENCES BETWEEN TREATMENT AND CONTROL MEANS ON ALL SUB-TASKS, AT MIDLINE

	Control Mean	Treatment Mean	Difference (T-C)
Overall			
Letters p/m	34,5	38,5	4,0
Di-/ Trigraphs p/m	15,2	17,9	2,7
CVCV Words p/m	13,4	15,1	1,8
Familiar Words p/m	9,4	11,1	1,7
Oral Reading Fluency	10,6	12,1	1,5
Reading Comp, I	4,4	4,7	0,3
Productive Listening	3,5	3,8	0,2
Vocabulary	5,3	5,2	-0,1
Phonemic Awareness	4,1	4,5	0,3
Grade 1			
Letters p/m	24,3	28,6	4,3
Di-/ Trigraphs p/m	6,2	8,5	2,3
CVCV Words p/m	6,3	7,5	1,2
Familiar Words p/m	4,2	5,4	1,2
Oral Reading Fluency	4,5	5,8	1,3
Reading Comp, I	2,3	2,6	0,3
Productive Listening	3,2	3,5	0,3
Vocabulary	5,1	5,0	-0,2
Phonemic Awareness	3,3	3,9	0,6
Expressive Vocabulary	9,1	9,6	0,4
Grade 2			
Letters p/m	44,8	48,1	3,3
Di-/ Trigraphs p/m	24,3	27,1	2,8
CVCV Words p/m	20,4	22,6	2,2
Familiar Words p/m	14,7	16,7	2,0
Oral Reading Fluency	16,7	18,2	1,6
Reading Comp, I	6,4	6,7	0,3
Productive Listening	3,9	4,1	0,2
Vocabulary	5,5	5,4	0,0
Phonemic Awareness	5,0	5,1	0,1
Oral Reading Fluency	15,3	16,3	1,0
Reading Comp, II	4,0	4,3	0,3
Sentence Choice	5,4	5,6	0,2

5.2 ESTIMATING TREATMENT EFFECTS

5.2.1 REGRESSION ANALYSIS OF INTERVENTION IMPACTS

The main equation to estimate is:

$$Y_{igsd1} = \beta_0 + \beta_1 (Treatment\ 1)_s + \boldsymbol{X}'_{i0}T + \mu_d + \varepsilon_{igsd1} \quad (I)$$

The outcome measure, Y_{igsd1} is a midline measure of reading proficiency for learner *i* in grade *g* of school *s*; *(Treatment)* is the dummy variable indicating the treatment status for school *s*; $\boldsymbol{X}'_{is0}$ is a vector of baseline controls, μ_d are the specific school district (or strata) fixed effects[58]; and ε_{igsd1} is the error term clustered at the school level. The parameter of interest (β_1) is the average treatment effect on learner reading outcomes.

Random assignment and the fact that the two groups were still balanced at midline ensures that a simple comparison of means across learners in the intervention and control schools provides unbiased estimates of the programme impacts. However, regression analysis of the programme impacts allows one to i) control for any incidental pre-randomisation 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 (but which are independent of treatment assignment). All results reported therefore control separately for each relevant measure of reading proficiency collected at baseline[59], learner level characteristics, household assets, as well as strata fixed effects[60].

In the cases where learners had missing data on a certain dimension of the baseline 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[61]. Since schools were randomly assigned to either the intervention or control groups, and not the individual learners, standard errors are clustered at the school level - the level of randomisation (Abadie, Athey, Imbens and Wooldridge, 2017)[62].

---

[58] Recall from section 4.3 that random assignment of schools occurred within each of the three the three educational districts.

[59] 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.

[60] Full lists of covariates reported in table 6 above.

[61] 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).

[62] Intuitively, two learners drawn at random from within the same school are more likely to have similar reading outcomes than two learners drawn randomly 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).

Heterogenous treatment effects are estimated as follows:

$$Y_{igsd1} = \beta_0 + \beta_1 (Treatment)_s + \beta_2 (Treatment \times x)_{is} + \boldsymbol{X}'_{i0}T + \mu_d + \varepsilon_{igsd1} \qquad (2)$$

Where $x_i$is the moderating variable of interest at the learner level (here the learner's baseline reading proficiency and/or their sex). The moderating variable is also included in the vector of baseline controls[63] .

5.2.2 INTERVENTION IMPACTS

5.2.2.1 EFFECTS MEASURED IN STANDARD DEVIATIONS

The results from the main estimation equation (1) are reported in figures 12 to 15 below. Estimated effect sizes are reported for standardised versions of the various sub-tasks, to provide 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. The exercise is repeated for each grade, including grade-specific controls[64]. 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 impacts.

The impact of the programme on reading proficiency for the full cohort of learners over one year of exposure is 0.17 s.d. (see the composite score estimate, figure 12). 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.

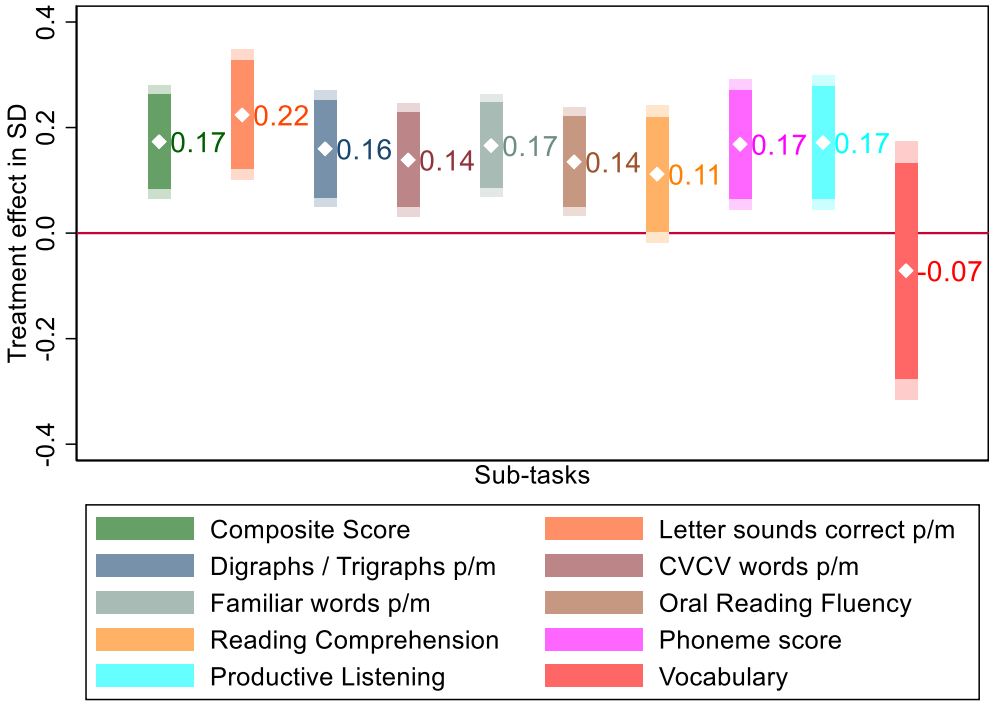
Figure 12 shows the estimated impacts for the full sample. The estimated impacts were relatively large and consistent across the range of emergent- and early literacy tasks. The largest point estimate of programme impact on any one sub-task was on the correct letter sounds per minute task (0.22 s.d.). 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.

[63] In the case of baseline reading proficiency, each constituent component of the composite score is controlled for separately – similar to the main estimation model.

[64] A large degree of the variation in midline composite reading proficiency is explained by the baseline covariates included in the models, driven by relatively high correlations between baseline and midline assessments. For the full sample. 67 percent of midline learner reading proficiency outcomes are explained by the baseline covariates (excluding treatment status). For Grade 2 specific models, baseline covariates explain an even higher 75 percent of midline variation. For Grade 1 reading proficiency, this is only 41 percent. A lot of precision is thus gained through the grade 2 learners specifically.

FIGURE 12: TREATMENT EFFECTS FOR COMMON TASKS ASSESSED ON BOTH GRADES, OVERALL AND BY SUB- TASK

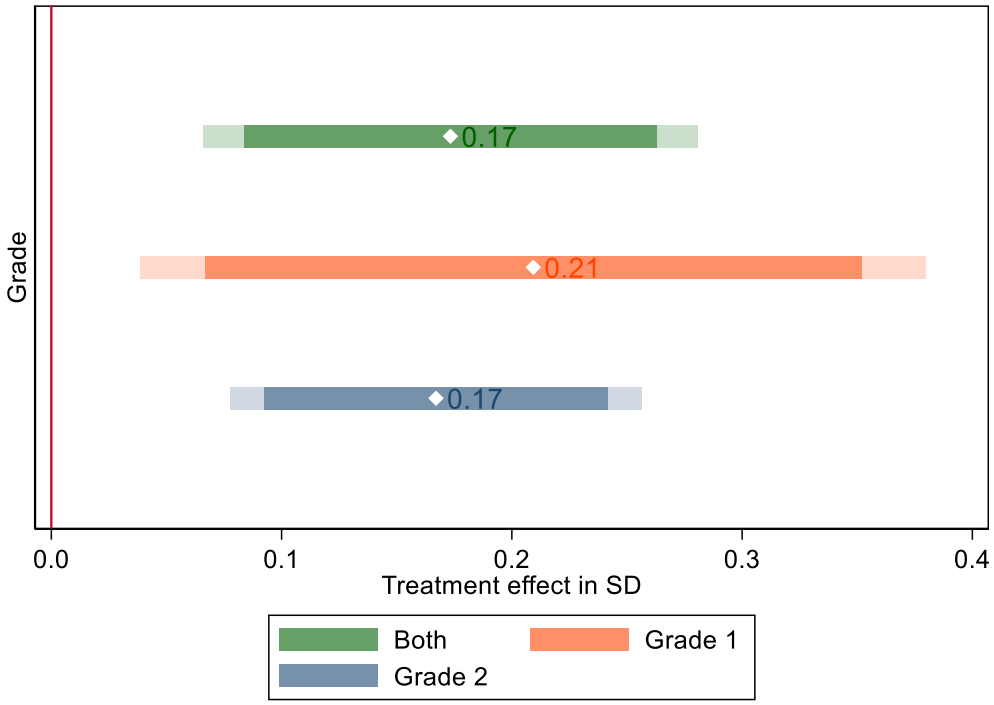


Reading comprehension, as the main outcome of interest of the programme, had a positive point estimate (0.11 s.d.) that was not statistically significant (see Figure 12 above, as well as the p-values reported in appendix Table A4). A smaller effect on reading comprehension at this stage accords with 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[65]. 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.

Figure 13 below displays the estimates 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[66]. Controlling for additional Grade 2 specific baseline tasks, the model for Grade 2 learners yields an effect size that is slightly lower at 0.16 s.d., but more precisely estimated. Given the overlapping confidence intervals, one cannot infer that the programme impact was larger on any one grade specifically.

[65] 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.
[66] 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 (refer to section 6.1.3).

FIGURE 13: TREATMENT EFFECTS ON READING PROFICIENCY, FOR FULL SAMPLE AND BY GRADE



Using the Grade 2 specific reading proficiency outcome measure, which weights reading fluency and comprehension skills more heavily, yields only a slightly lower estimate of the intervention's impact: 0.15 s.d (see Grade 2 specific Figure 14 below). Figure 14 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 productive listening is comparatively large (0.15 s.d.), but the effects on these two tasks are not statistically distinguishable from zero.

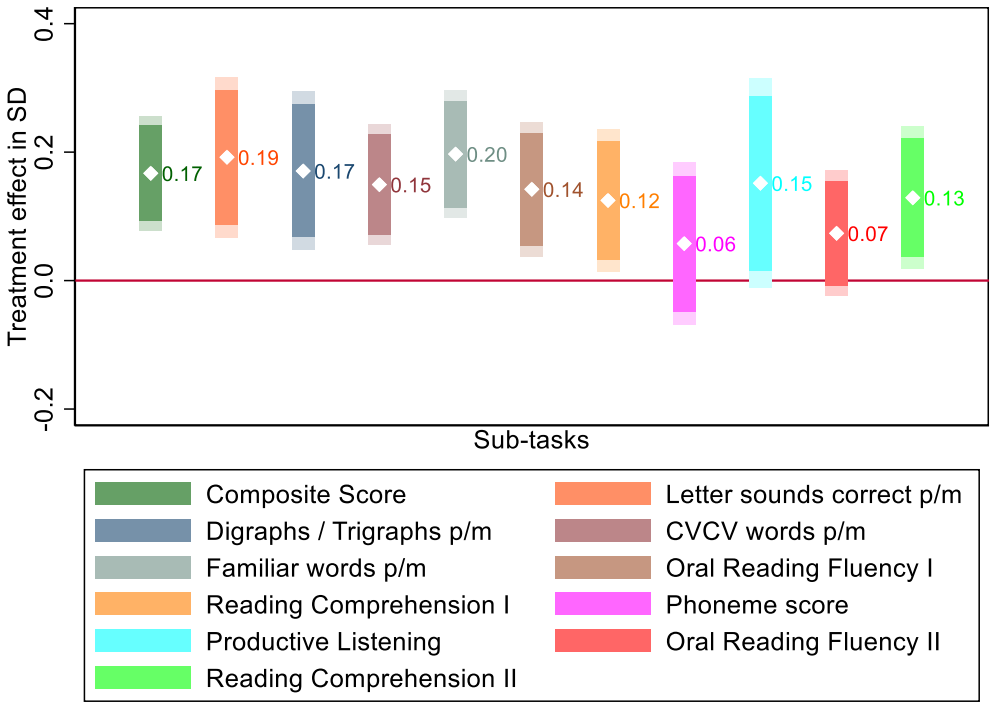
Grade 2 learners' reading fluency was assessed on two separate passages at midline. The first passage was the same passage used to assess reading fluency and -comprehension for Grade 2 learners at baseline, whilst the newly introduced second passage was slightly longer and more challenging[67]. The effect 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 of the programme on reading

[67] 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.

comprehension for both passages. Grade 2 learners in treatment schools score around 0.12 to 0.13 s.d. higher than 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[68]. Considered overall, the programme significantly shifts reading comprehension outcomes at the Grade 2 level.

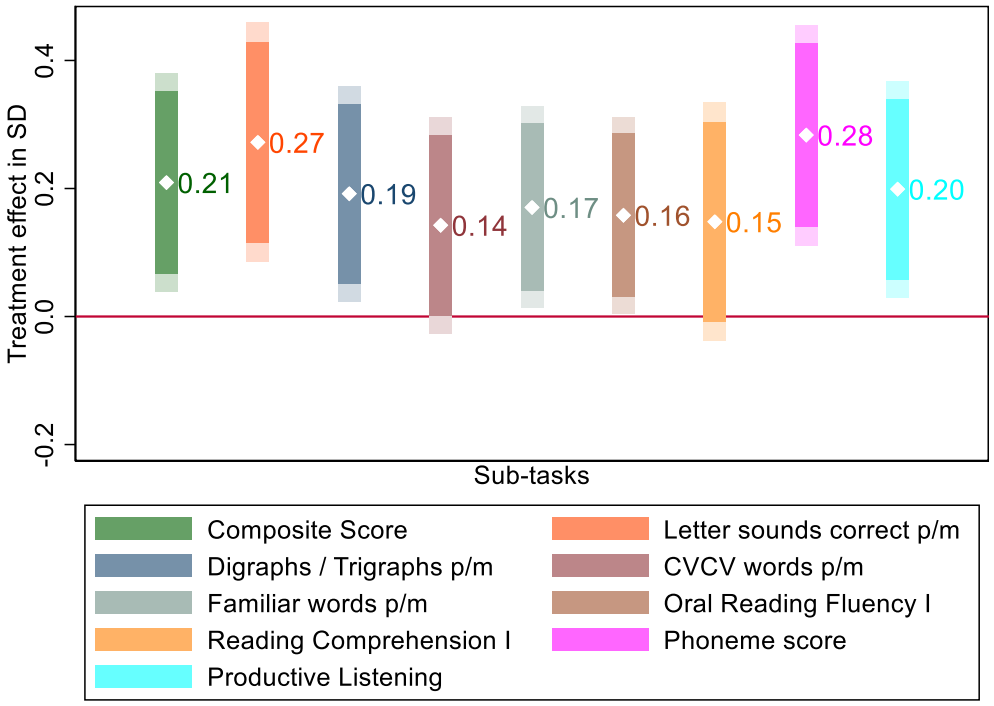
FIGURE 14: TREATMENT EFFECTS FOR GRADE 2'S ONLY, OVERALL AND BY SUB-TASK



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 15). There is a 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).

FIGURE 15: TREATMENT EFFECTS FOR GRADE 1 SAMPLE, OVERALL AND BY SUB-TASK

[68]As discussed in annex IV in the appendix, the task seems be limited in its ability to differentiate among learners' underlying reading comprehension ability.



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[69]. 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. 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[70].

5.2.2.2 *Effects measured in terms of a year of learning*

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 provide 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 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

[69] 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.

[70] 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%.

measurement requires a sub-task to have been assessed on the relevant grade at both the baseline and midline assessments[71].

For each sub-task conducted on Grade 2 learners, Table 10 below displays for the control group the mean outcome at baseline, the standard deviations of baseline outcomes, the growth (or difference in means) between baseline and midline, and the number of observations on which the estimates are based. The second part of the table provides the estimate of the effect size (also in task-specific units, like words read correctly per minute) and reinterprets this as a percentage of the learning that took place 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 correct letter sounds per minute[72]. This grew to 45 letter sounds per minute at midline (see the third column in Table 10). The estimated impact of four additional correct letter sounds per minute (fifth column, Table 10) for Grade 2 learners in the intervention schools equates to 27 percent of a year's worth of learning (roughly one school term). 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 control schools.

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 percent higher comprehension scores by intervention school Grade 2 learners translates to 24 percent of a year of learning. This gives practical significance to the 0.14 s.d. effect size on reading comprehension from Figure 14. Overall, reading comprehension gains for Grade 2's in intervention schools are promising and large relative to status quo levels of learning.

TABLE 10: 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%

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

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

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 I	4,3	4,6	2,1	278	0,5	24%

For Grade 1 learners, only four of the tasks that were assessed at both baseline and midline had significant effect sizes (Table 11). 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.

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, the average digraphs and trigraphs correctly identified per minute moved from a base of almost zero to six at midline for control school learners[73]. 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. In absolute terms, the effect size is similar to the effect of the intervention on Grade 2 learners (three to four letter sounds). However, it is almost three times larger when viewed relative to the amount of learning that occurred in the Grade 1 comparison group.

For the phonemic awareness task, control school Grade 1's basically doubled their scores 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 point treatment effect translates to 46 percent of a year of learning. Considered collectively, the larger effects on the four foundational literacy skills in Table 11 for Grade 1 intervention school learners (relative to intervention group Grade 2's) provide further suggestive evidence that literacy skill acquisition is 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.

[73] At baseline, 96 percent of all Grade 1 learners could not identify a single digraph or trigraph. 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 11: 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%

5.3 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 by learner (equation (2)). Results from other structured pedagogical programmes similar to the Funda Wande intervention (like the EGRS 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, two learner level characteristics are investigated for any differential treatment effects at this stage of the evaluation[74]: i) baseline reading proficiency and ii) gender.

5.3.1.1 IMPACTS BY INITIAL READING PROFICIENCY

[74] The risk to investigating multiple possible sources of differential treatment effects is called data mining. In other words, if one tests for differential treatment effects by a whole range of characteristics and sub-combination of them, its 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, one 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.

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[75].

Looking back to Figure 8 and 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[76]. Figures 16 and 17 below 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 16 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[77].

Another way to read the graph is to compare the number of 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) than learners from control schools. At the median 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).

[75] 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.

[76] 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.

[77] 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.

FIGURE 16: DISTRIBUTIONS OF MIDLINE LETTER SOUNDS BY TREATMENT STATUS – GRADE 1

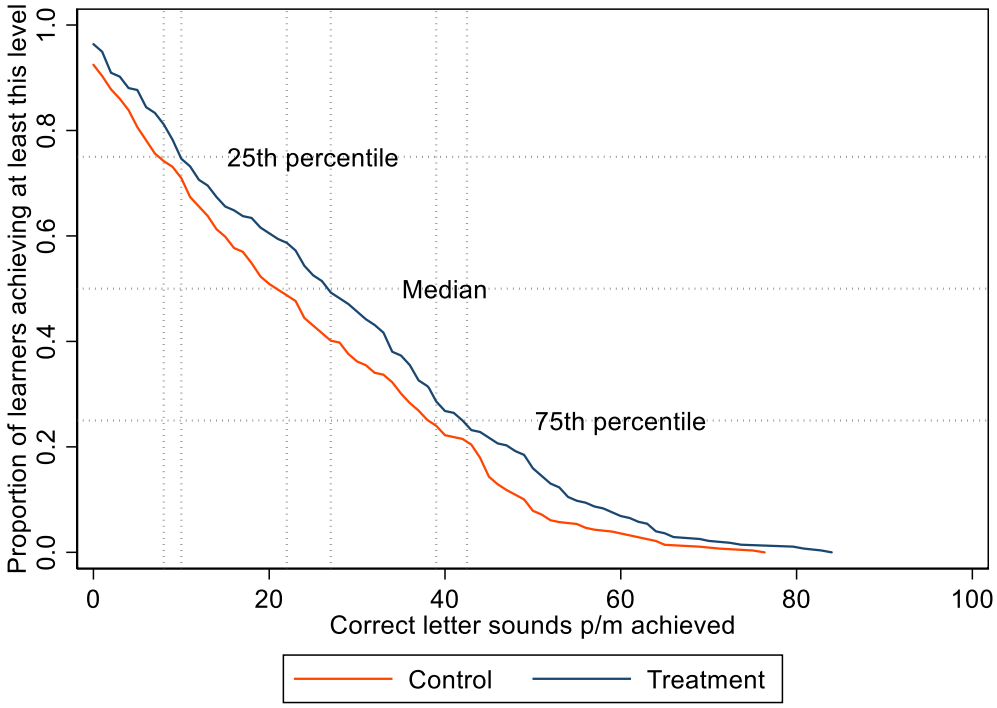
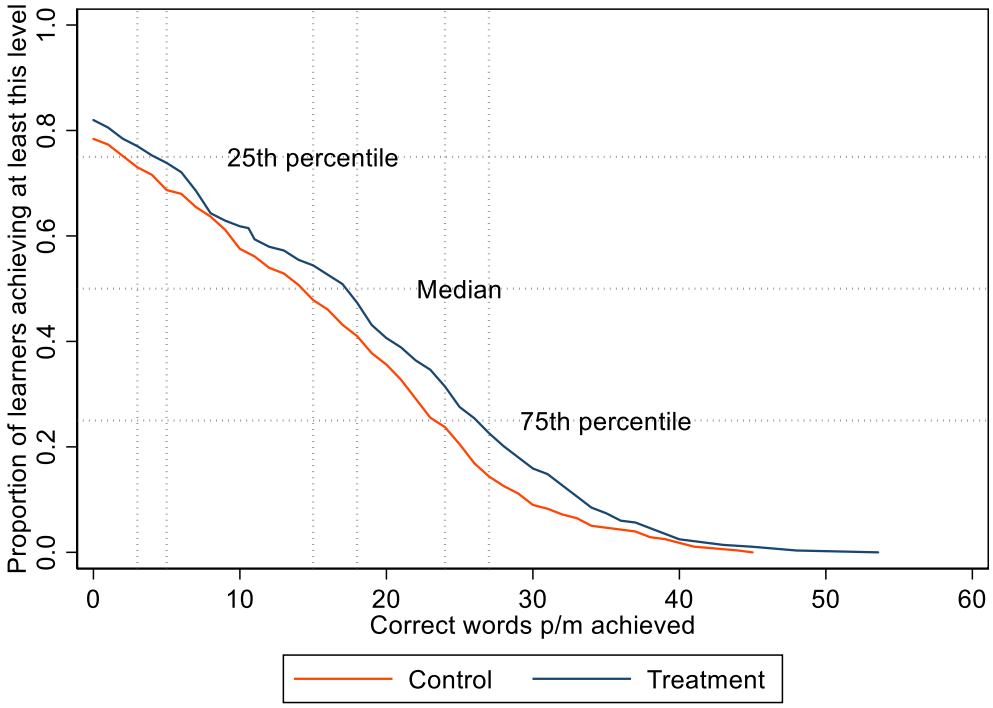


Figure 17 below performs the same exercise, but for the distribution of Grade 2 learners' correct familiar words read per minute scores, for the two groups respectively. Again, learners in the intervention group outperform control school Grade 2's across the entire distribution. At the lower end of the 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 per minute (five, as opposed to three in the control group). At both the medians and the 75th percentiles, intervention school learners read 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 mid-to-high range of the distribution.

FIGURE 17: DISTRIBUTIONS OF MIDLINE FAMILIAR WORDS BY TREATMENT STATUS – GRADE 2



To investigate whether the impact of the programme is statistically significantly different depending on learner's baseline reading proficiency, the same models as in the main analysis are used, 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 12 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.

The first column in table 12 suggests that there is not a 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 two tests for a quadratic relationship between the intervention and learners baseline reading proficiency[78]. 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.

TABLE 12: INTERVENTION EFFECTS BY BASELINE READING PROFICIENCY

[78] The model in column (2) of Table 12 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.

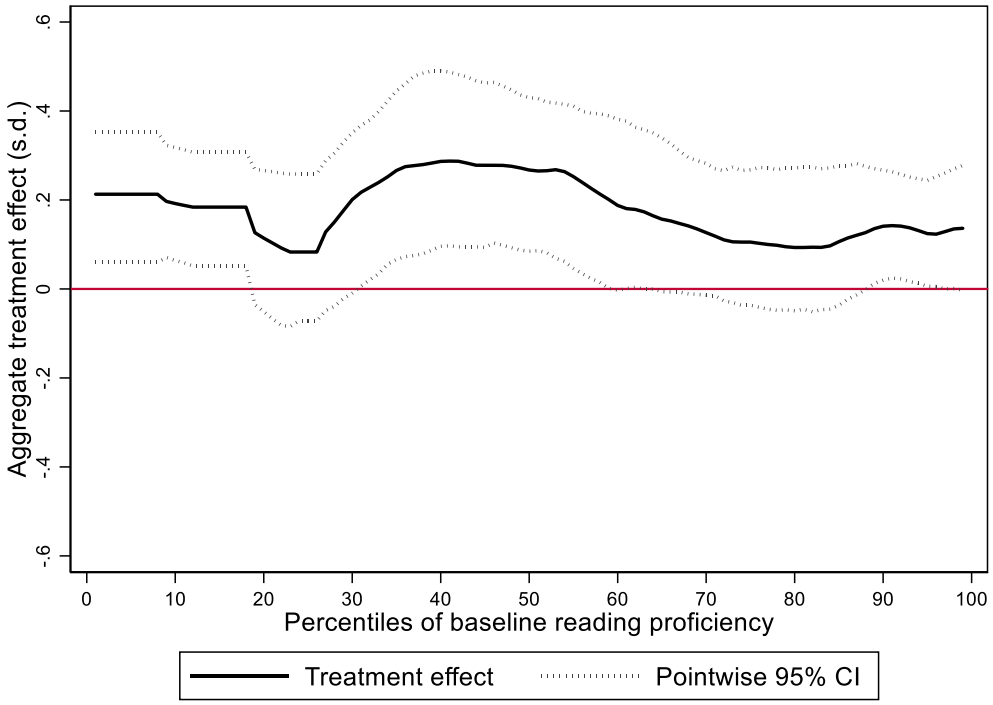
	Outcome variable: Midline Composite score	
Treatment	0.174***	0.181***
	(0.0538)	(0.0648)
Treatment x baseline composite score	-0.0420	-0.0320
	(0.0440)	(0.0508)
Treatment x baseline composite score squared		-0.00711
		(0.0257)
Sample	FULL	FULL
Observations	1,104	1,104
R-squared	0.678	0.684
Heterogeneous treatment effect: P-value	0.344	0.593

Note: Standard errors reported in brackets.

Figure 18 displays the estimated programme impact at each percentile of learner baseline reading proficiency[79]. 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 earlier comparisons of the distributions of midline reading proficiency scores between treatment and groups.

[79] More precisely, Figure 18 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 18: NONPARAMETRIC INTERVENTION IMPACTS BY BASELINE READING PROFICIENCY



One might also be interested in whether the programme's effects are as consistent based on learners' 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 learners 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 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 19 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[80]. The intervention was effective in shifting the reading outcomes of those learners who

[80] 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

were the least proficient in their class at baseline (effect size of 0.24 s.d.) and for the most proficient readers (0.26 s.d.). There is not a significant difference in impact on the most or least proficient readers.

FIGURE 19: INTERVENTION IMPACTS ESTIMATED ON MOST- AND LEAST PROFICIENT READERS AT BASELINE, RESPECTIVELY

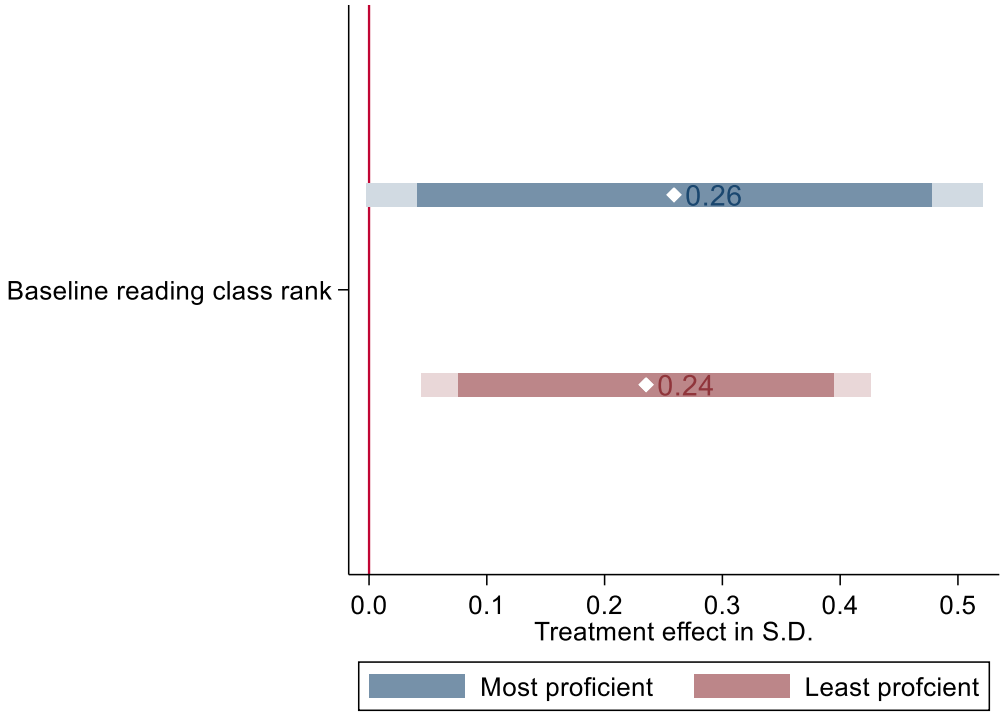
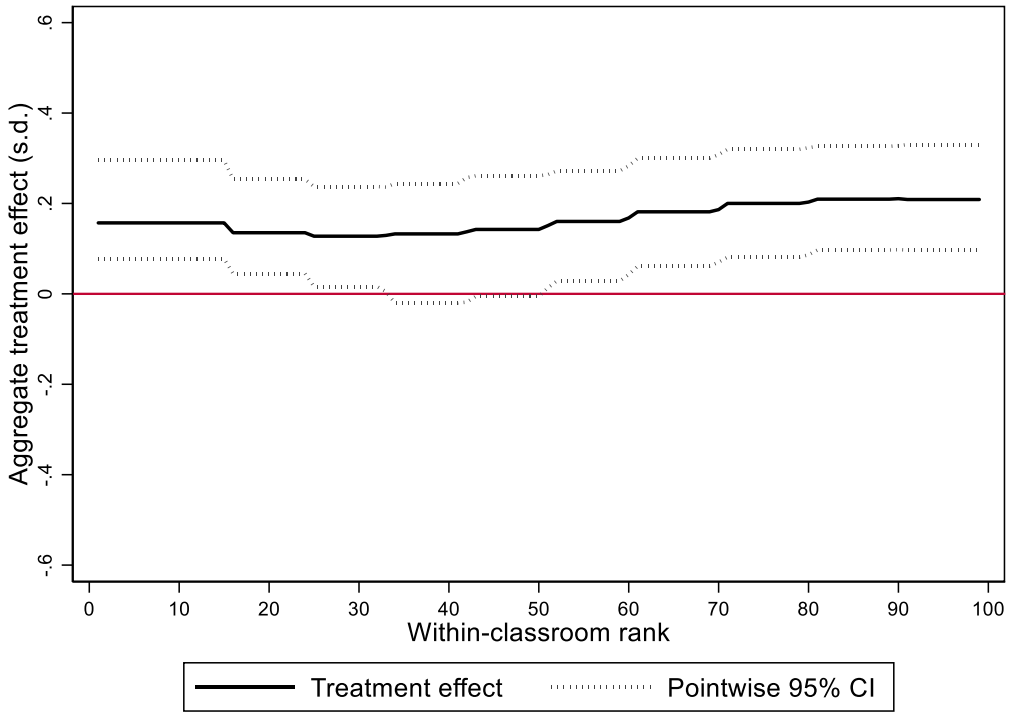


Figure 20 repeats the same non-parametric estimation of treatment impacts based on learners' baseline reading proficiency as in figure 18 above, but now based on a learner's baseline rank within the ten learners assessed per class. Treatment effects are consistent and positive, independent of where a learner ranked within their classroom at baseline. Together, these results suggest that neither learners' 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.

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 20: NONPARAMETRIC INTERVENTION IMPACTS BY WITHIN CLASS RANK OF BASELINE READING PROFICIENCY



5.3.1.2 GENDER

For both grades, there is no differential impact between boys and girls overall. Table 13 tests whether the impact of the programme is significantly different depending on a learner's gender by including an interaction term for intervention group status and whether or not a learner is female. Column one below shows that while the point estimate of the 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.

However, there is stronger evidence of differential treatment impacts by gender when the analysis is redone on each grade separately. Column two in Table 13 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 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 point estimate for Grade 1 girls (though Grade 1 point estimates are less precisely estimated and the differential impact for Grade 1's by gender thus not significant and uncertain).

TABLE 13: INTERVENTION EFFECTS BY GENDER

	Full	Grade 1	Grade 2
	Outcome variable: Midline Composite score		
Treatment	0.199***	0.145	0.269***
	(0.0664)	(0.0900)	(0.0693)

Treatment x female	-0.0504 (0.0755)	0.128 (0.0941)	-0.194* (0.0984)
Observations	1,104	552	552
R-squared	0.678	0.428	0.756
Heterogeneous treatment effect: P-value	0.507	0.178	0.0541

Note: Standard errors reported in brackets.

Graphically, figures 21 and 22 provide further support that the intervention has a differential impact on boys and girls, depending on where they are in their developmental trajectories. The figures provide a comparison of the distributions of midline reading proficiency scores for the intervention and control groups, by gender for each grade separately.

Grade 1 girls outperform their boy counterparts across the board and also seen to benefit more from the programme (figure 21). In Grade 2, boys in the intervention schools catch up with their girl counterparts (even though boys in both the intervention and control groups still score significantly lower than their girl counterparts, in both control and intervention schools – figure 22).

Whilst these results are only suggestive at this stage, and confirmatory findings in subsequent round of analysis and/or larger sample sizes would be required, this suggests that by Grade 2 boys are more likely to have reached a stage in their development to benefit from the intervention. Grade 2 girls, in contrast, see only marginal if any benefit from the programme (based on the foundational literacy skills measured, at least). As we see the largest effects for the foundational reading skills in Grade 1 (for e.g. letter sounds and phonemic awareness), it is possible that improvements in the higher order skills take longer to manifest or happen in smaller increments. In Grade 2, more of the female learners are in this zone and therefore measured effects may appear more muted.

FIGURE 21. COMPOSITE READING PROFICIENCY BY INTERVENTION STATUS AND GENDER – GRADE 1

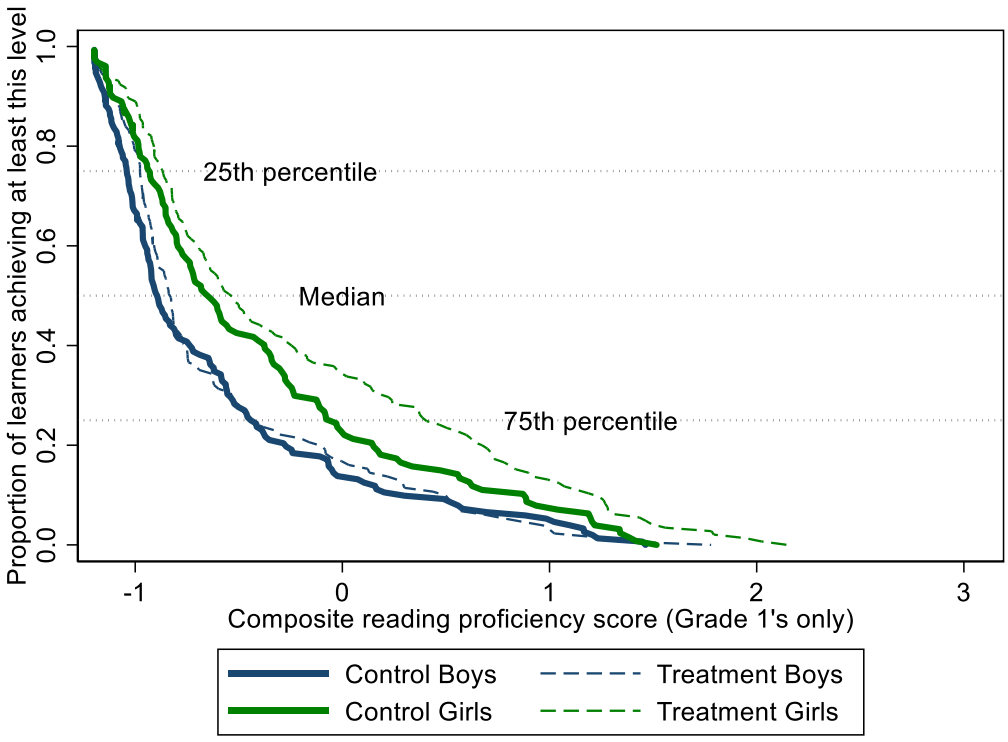
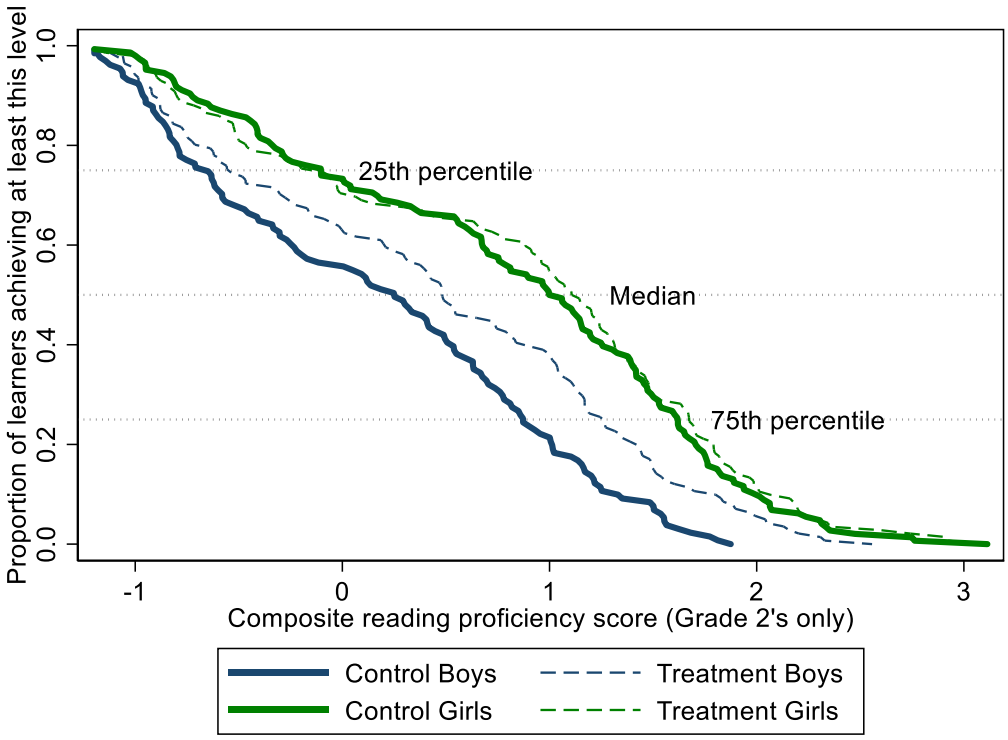
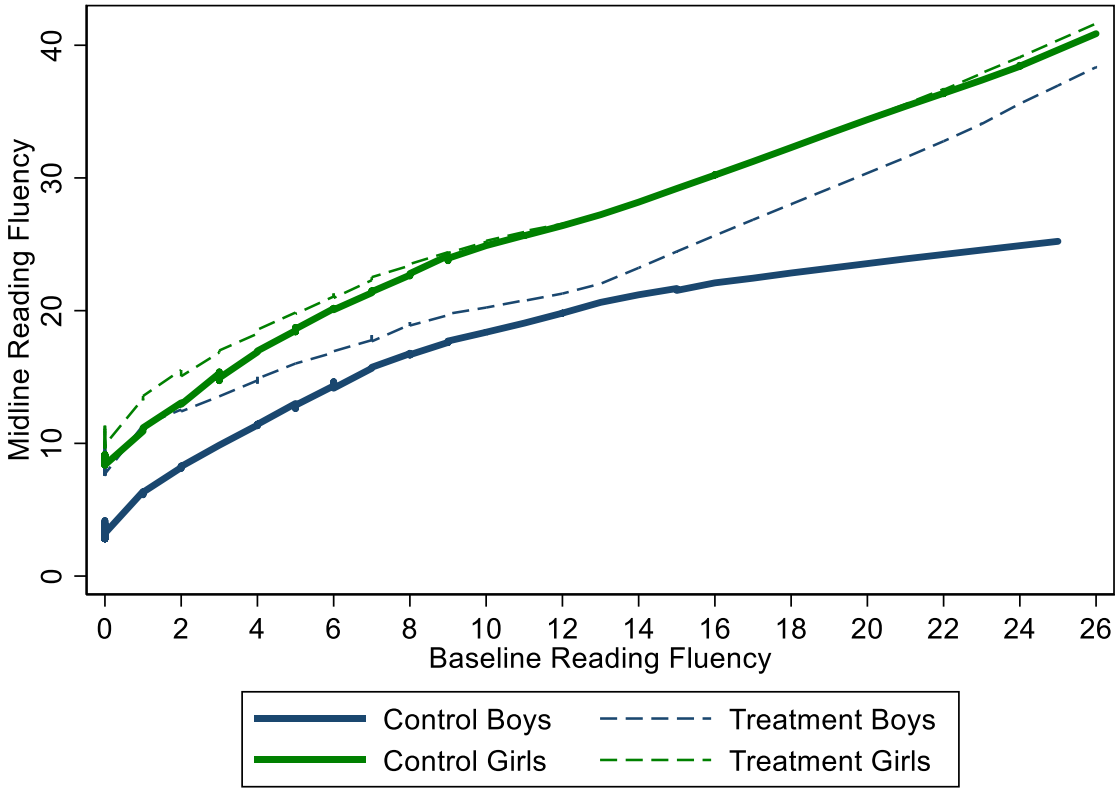


FIGURE 22: COMPOSITE READING PROFICIENCY BY INTERVENTION STATUS AND GENDER – GRADE 2



Digging deeper into the differential treatment effects for boys and girls in Grade 2, Figure 23 suggests that boys at higher levels of baseline reading fluency experience much larger gains (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 intervention school girls with 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 baseline correct words per minute and up). 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 23: GREATER GAINS IN READING FLUENCY FOR BEST PERFORMING BOYS AT BASELINE (GRADE 2)



Altogether, these results suggest that the programme has the greatest impact on learners at a certain stage of their literacy development trajectories (and that the latter differs systematically by gender). 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, given the right educational inputs. Boys, in turn, who are generally lagging behind their female counterparts, are more likely to reach such a point where they have the potential for rapid early reading skill acquisition only somewhat later - in Grade 2.

6 INTERMEDIATE OUTCOMES: TEACHERS

The Funda Wande intervention specifically targets teachers, given that pedagogy in the majority of South African classrooms is largely communalized (Hoadley 2019). This section explores 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. It also examines whether there are any shifts towards more individualised forms of learning through teachers reports on their use of the graded readers provided.

6.1 FORMATIVE ASSESSMENT AND AWARENESS OF LEARNERS' ACTUAL READING LEVELS

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 teacher ranked as the most/least proficient reader, one can therefore compare where they actually

ranked out of the learners assessed based on an objective measure of reading proficiency. 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 those learners who are truly the most/least proficient readers 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[81]. All results are thus only suggestive at this stage.

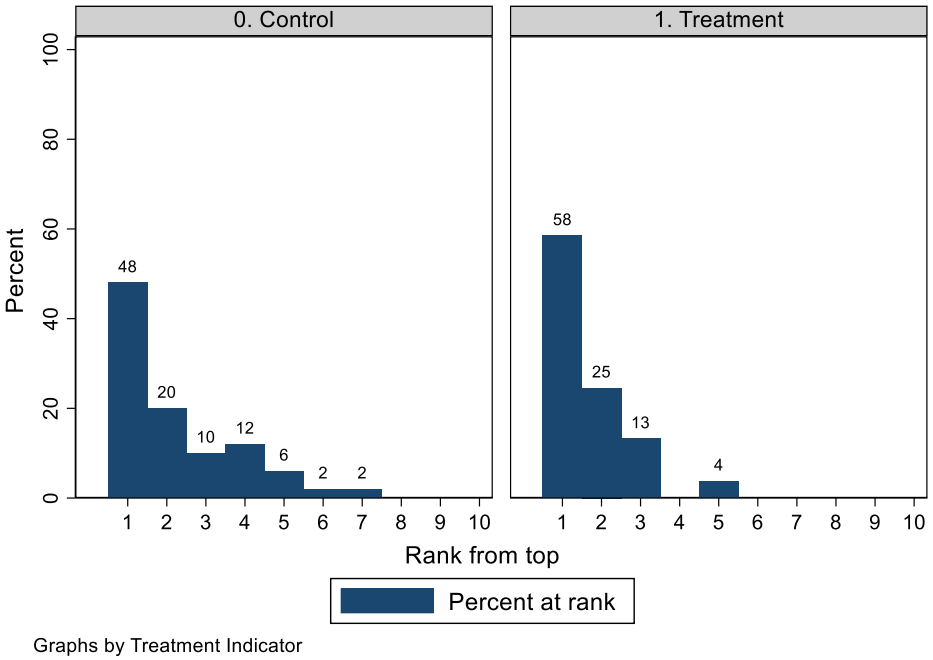
In terms of measurement, no exact definition or criteria on how to determine the most/least proficient readers in the list was provided to the teachers. An obvious candidate would be oral reading fluency. However, 55 percent of Grade 1's[82] scored zero on the reading fluency task and one assumes that 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. The midline composite reading proficiency measure is used instead, as this better discriminates at the bottom of the distribution. Learners' actual rank based on this composite measure of proficiency is used for the comparison of most proficient" and "least proficient" identified readers actually performed by treatment status[83].

FIGURE 24: ACTUAL RANK AMONG ASSESSED CLASSMATES OF TEACHER IDENTIFIED "MOST PROFICIENT READER"

[81] 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.

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

[83] The class rank used to determine the most and least proficient readers follows the same ranking methodology as outlined in section 5.3.1.1 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.



When comparing the accuracy of the rank between teachers in control and treatment schools, teachers in the intervention schools were more accurate on average in their predictions of the most proficient reader. Figure 24 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. An ordered logit regression was used to tested whether treatment teachers were better able to identify the most proficient readers in their class[84]. 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 A6).

Results are less clear on teachers' ability to identify the "least proficient reader" of those assessed in their class. Here 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 25, a greater proportion of the "least proficient readers", as identified by control school teachers, 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", a formal test is conducted using an ordered logit. The point estimate is negative suggesting

[84] Moe precisely, a regression of the actual rank of the learner that the teacher identified as most proficient on an indicator that the teacher was in an intervention school. The regression included strata fixed effects and standard errors were clustered at the level of the school.

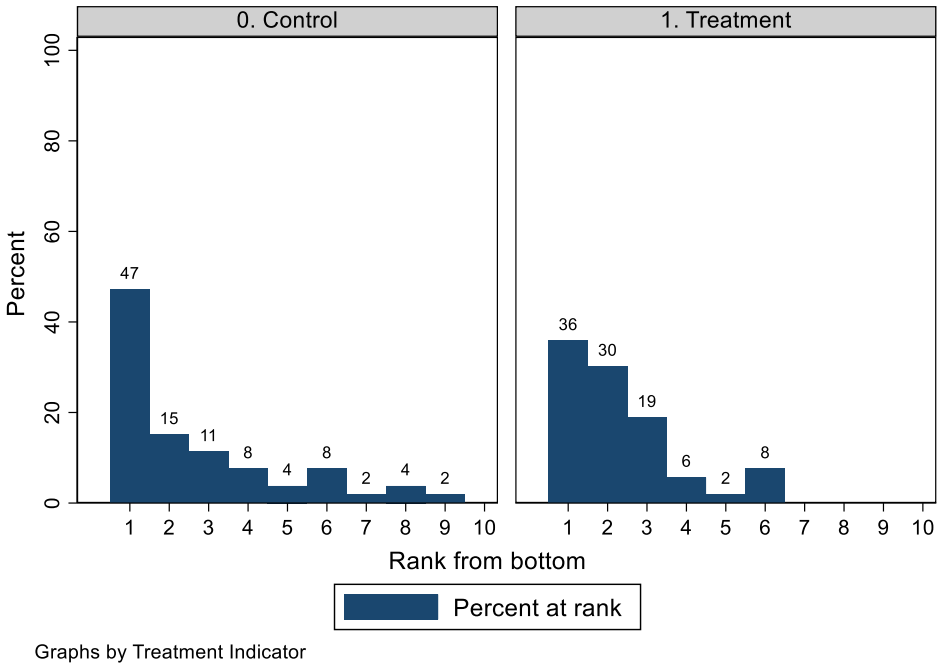
that treatment teachers are better able to identify poorer performing learners, but the estimate is not statistically significant (output shown in appendix Table A6)[85].

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.

FIGURE 25: ACTUAL RANK AMONG ASSESSED CLASSMATES OF TEACHER IDENTIFIED "LEAST PROFICIENT READERS"

[85] Similar analyses were conducted 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.



Similarly, further exploratory analysis reported in Annex V further suggests that intervention school teachers are better attuned to the actual reading levels of the learners in their class. These results focus specifically on teachers ability to gauge the average reading performance of their class on average, and not only in picking out the most- and least proficient readers.

6.2 USE OF GRADED READERS

In 2019 for the first time, the ECDoE distributed a grade-level anthologies 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 the appendix of the midline report on the Funda Wande intervention (Ardington and Meiring, 2020). In summary, distribution was almost universal, with 93 percent of principals/HODs reporting that anthologies were delivered to their school. The vast majority (96%) of teachers report receiving copies, with only 1 percent indicating that the copies received were not enough for each learner to have their own copy. Most learners reported receiving their own copy of the anthology (91 percent) and that they were allowed to take the anthologies home (92 percent).

Given that teachers in both control and intervention schools were furnished with the exact same resources, it provides an opportunity to investigate how teachers in the two groups put the newly introduced resources into use in the classroom. Figure 26 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 sixteen percent of control teachers. Intervention teachers also report using more of the stories in the anthology than control teachers (Figure 27).

FIGURE 26: FREQUENCY OF USE OF VULA BULA STORIES IN CLASS – TEACHER REPORT

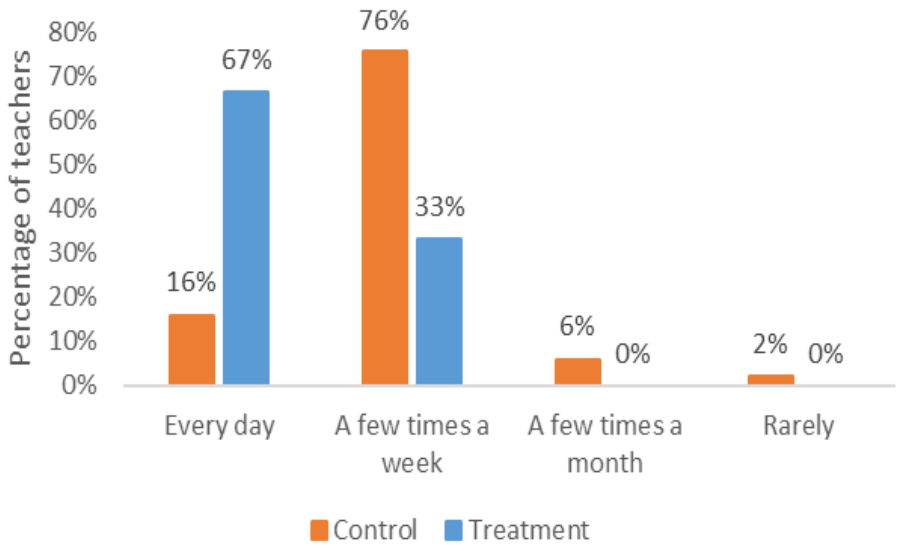
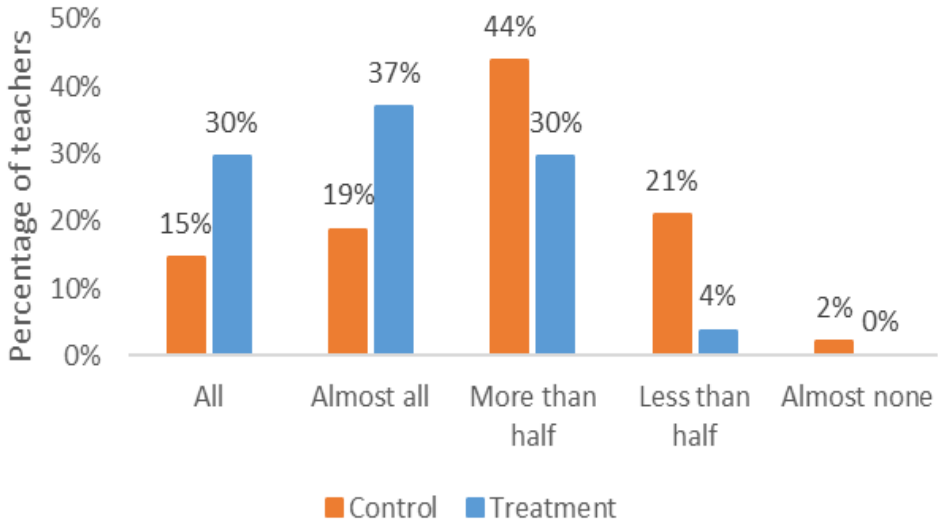


FIGURE 27: NUMBER OF VULA BULA STORIES USED – TEACHER REPORT



Learners were also asked how often the anthologies were used in the class (Figure 28) and how many of the stories they had read (either themselves or by an adult) (Figure 29). Learner reports of frequency of use and number of 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[86].

FIGURE 28: FREQUENCY OF USE OF VULA BULA STORIES IN CLASS – LEARNER REPORT

[86] 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).

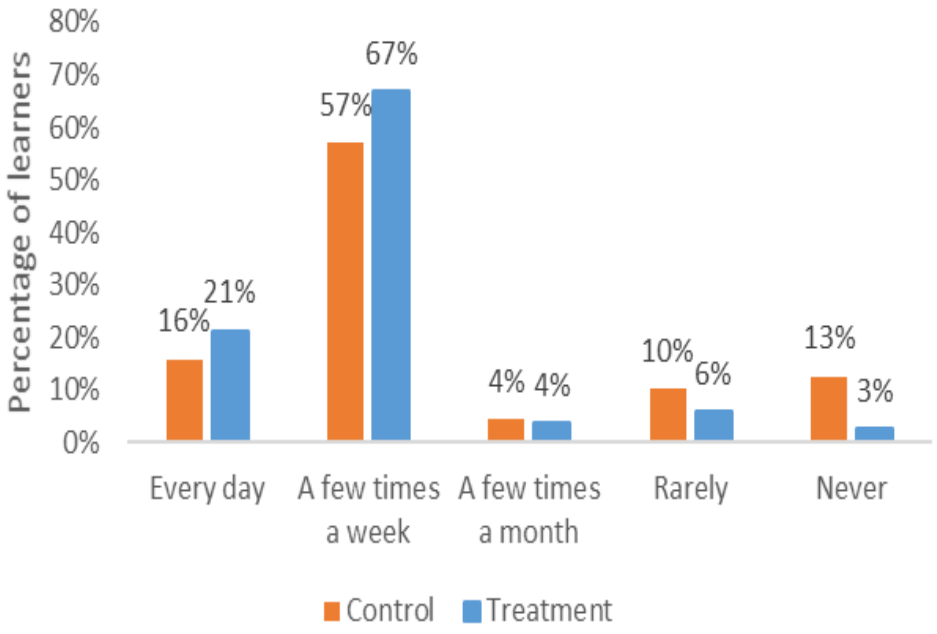


FIGURE 29: NUMBER OF VULA BULA STORIES USED – LEARNER REPORT

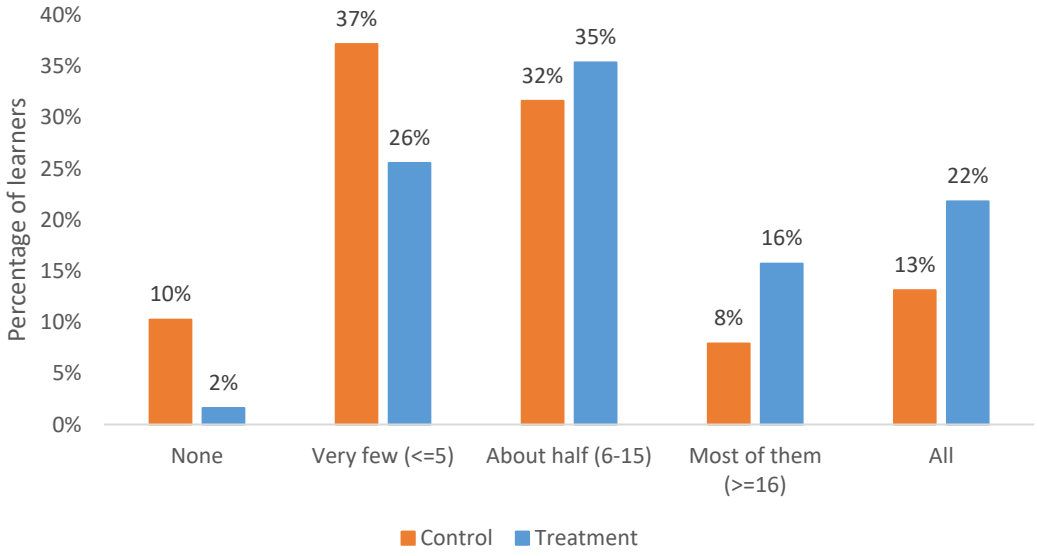


Table 14 shows the range of activities that the teachers conduct with the anthologies by intervention status. In almost half of both intervention and control classrooms, the teacher reads aloud from the anthology while learners listen. The other form of communalised learning, the whole class reading 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 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 explanation for the differences in use presented in table 14.

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%

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 intervention teachers. The in-depth qualitative classroom observations planned for the second year of the intervention will shed more light on whether and how Funda Wande is shifting reading pedagogy in the classroom.

7 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 sub-tasks that were reliably measured, 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, 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 impact on Grade 1 learners' foundational skills (letter sound recognition and phonemic awareness) is 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. This is 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, there is only 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 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 EGRS studies are an important comparator and benchmark for the Funda Wande intervention. From a research perspective it is important to build on previous research insights, assessing the generalizability of this type (or 'class') of intervention and understanding why and how they are different in design and/or effectiveness. It is also important from a policy perspective, as the EGRS studies represent a culmination of the South African government's evidence base and thinking on the most effective ways of improving teacher capacity and early learning outcomes at a national, public schooling system level. For ease of comparison throughout, the details and results of the EGRS studies are therefore summarized in annex I.

The midline results from the EGRS 1 study (Cilliers et al., 2019, Taylor et al., 2016) 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, 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)[87]. However, 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. Annex 1 provides a more in-depth comparison of the Funda Wande and EGRS midline results.

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.

[87] 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.

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[88] (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 annex III). These sample criteria guided the process of selecting the 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 learner 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[89] . If effect sizes therefore hold with some consistency across provinces, for schools with different languages of instruction, both rural/urban

[88] 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).

[89] 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).

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 Cape) and in more than one language of instruction (Setswana Home language, English as first additional language and isiXhosa Home language).

Given the consistent positive impacts found for the structured pedagogy programmes that have been assessed in the South African context, these programmes are arguably past the proof of concept stage and have shown some degree of consistency in shifting learner reading outcomes across these different contexts. 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. These questions provide fruitful avenues for future research.

8 REFERENCES

Abadie, A., S. Athey, G.W. Imbens, and J. Wooldridge (2017). 'When Should You Adjust Standard Errors for Clustering?'. NBER Working Paper 24003. Cambridge, MA: National Bureau of Economic Research (NBER). Available at: https://www.nber.org/papers/w24003.pdf (accessed 7 February 2019).

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

Allington, R.L., and P. Johnston. (2002). *Reading to learn: Lessons from exemplary 4th grade classrooms*. New York: Guilford.

Allington, R.L., and P. Johnston. (2001). 'What do we know about effective fourth grade teachers and their classrooms?' *In* Roller, C. (*Ed*.). *Learning to teach reading: Setting the research agenda*. Newark, DE: International Reading Association. pp. 150-165.

Altinok, N. (2013). 'The impact of teacher knowledge on student achievement in 14 sub-saharan African countries'. Background paper prepared for the Education for All Global Monitoring Report. Paris: UNESCO. Available at: http://unesdoc.unesco.org/. (accessed 15 May 2019).

Angrist, J. and Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Ardington, C. (2019). 'Impact Evaluation of the Funda Wande Coaching Intervention: Baseline Findings'. Cape Town: SALDRU, University of Cape Town.

Ardington, C., and C.J. Meiring (2020). 'Midline I: Impact Evaluation of the Funda Wande Coaching Intervention'. Cape Town: SALDRU, University of Cape Town. Available at: https://fundawande.org/news/impact-evaluation-of-funda-wande-coaching-intervention-midline-findings-23 (accessed 7 July 2020).

Athey, S., and G.W. Imbens (2016). 'The Econometrics of Randomized Experiments'. In A.V. Banerjee and E. Duflo (eds), *Handbook of Economic Field Experiments: Volume 1* (pp 73–140). Cambridge, MA: Elsevier.

Ball, J. (2010). *Enhancing learning of children from diverse language backgrounds: mother tongue-based bilingual or multilingual education in the early years'* Paris: UNESCO.Bates, M. A., and R. Glennerster. (2017). 'The generalizability puzzle'. *Stanford Social Innovation Review, Summer,* 50–54.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shortland, and M. Walton (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 J. Das (2020). 'Teacher Value-Added in a Low-Income Country'. *American Economic Review: Economic Policy,* 12(1): 62–96.

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

Bold, T., D. Filmer, G. Martin, E. Molina, C. Rockmore, J. Svensson, and W. Wane (2017). 'What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa'. World Bank Policy Research Working Paper 7956. Washington, DC: World Bank.

Brunette, T., B. Piper, R. Jordan, S. King, and R. Nabacwa (2019). 'The Impact of Mother Tongue Reading Instruction in Twelve Ugandan Languages and the Role of Language Complexity, Socioeconomic Factor, and Program Implementation'. *Comparative Education Review*, 63(4): 591–612.

Bruns, B., and J. Luque (2014). 'Great Teachers: How to Raise Student Learning in Latin America and the Caribbean'. Washington, DC: World Bank.

Bruhn, M., and D. McKenzie (2009). 'In Pursuit of Balance: Randomization in Practice in Development Field Experiments'. *American Economic Journal: Applied Economics*, 1(4): 200–32.

Carnoy, M., L. Chisholm, and B. Chilisa (2012). *The Low Achievement Trap: Comparing Schooling in Botswana and South Africa*. Pretoria: HSRC Press.

Chauraya, M., Brodie, K. (2017). 'Learning in professional learning communities: shifts in mathematics teachers' practices'. *Afr. J. Res. Math. Sci. Technol. Educ*. 21 (3): 223–233.

Chetty, R., J. Friedman, and J. Rockoff (2014). 'Measuring the Impacts of Teachers II: Teacher Value Added and Student Outcomes in Adulthood'. *American Economic Review*, 104(9): 2633–79.

Cilliers, J., B. Fleisch, C. Prinsloo and S. Taylor (2019). 'How to Improve Teaching Practice? An Experimental Comparison of Centralized training and in-classroom coaching'. *Journal of Human Resources*, 55: 926–62.

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?' Gui^{2}de working paper series no. 6. Georgetown University Initiative on Innovation, Development and Evaluation (Gui^{2}de). Available at: https://repository.library.georgetown.edu/handle/10822/1055275. (accessed 15 October 2019).

City, E.A., R.F. Elmore, S.E. Fiarman, and L. Teitel (2009). *Instructional Rounds in Education: A Network Approach to Improving Teaching and Learning*. Cambridge, MA: Harvard Education Press.

Conn, K.M. (2017). 'Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations'. *Review of Educational Research*, 87(5): 863–98.

Cummins, J. (2007). 'Rethinking Monolingual Instructional Strategies in Multilingual Classrooms.' *Canadian Journal of Applied Linguistics,* 10 (2): 221–40.

Cummins, J. (2001). *Negotiating Identities: Education for Empowerment in a Diverse Society*. 2nd ed. Los Angeles: California Association for Bilingual Education.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.

DBE (2018). *School Monitoring Survey 2017/18: Quantitative and Qualitative Summary* Report. Pretoria: Department of Basic Education

DBE. (2016). Report of the ministerial task team appointed Minister Ange Motshekga to Investigate Allegations into selling of posts of educators by members of teachers' unions and department official in provincial education departments. Pretoria.

DBE. (2015). Action plan to 2019: Towards the realisation of schooling 2030. Pretoria: government Printer.

DBE. (2014). Second Detailed Indicator Report for Basic Education Sector. Pretoria.

Deaton, A. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, *48*(2), 424–45.

Evans, D.K., and A. Popova (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): 242–70.

Fleisch, B., V. Schöer, G. Roberts, and A. Thornton (2016). 'System-Wide Improvement of Early-Grade Mathematics: New Evidence from the Gauteng Primary Language and Mathematics Strategy'. *International Journal of Educational Development*, 49: 157–74.

Flesich, B. and Schoer, V. (2014) 'Large-scale instructional reform in the global South: insights from the mid-point evaluation of the Gauteng Primary Language and Mathematics Strategy'. *South African Journal of Education*, 34(3): 1 -11.

Fleisch, B., S. Taylor, V. Schöer, and T. Mabogoane (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.

Fullan, M. (2007). *The New Meaning of Educational Change*. 3rd ed. London: Routledge.

Fullan, M., Knight, J. (2011). 'Coaches as system leaders'. *Educational Leadership*, 69 (2): 50–53.

Funda Wande (2018). Funda Wande: Reading for Meaning [Project Document]. Cape Town: Funda Wande.

Ganimian, A.J. and Murnane, R.J. (2016). 'Improving Educational Outcomes in Developing Countires: Lessons from Rigourous Impact Evaluations'. NBER Working Paper 20284. Cambridge, MA: National Bureau of Economic Research (NBER). Available at: http://www.nber.org/papers/w20284 (accessed 7 February 2019).

Glennerster, R., and Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide*. Princeton; Oxford: Princeton University Press

Glewwe, P., E.A. Hanushek, S. Humpage, and R. Ravina (2014). 'School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010'. In P. Glewwe (ed.). *Education Policy in Developing Countries* (pp. 13–64). Chicago, IL: University of Chicago Press.

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

Glewwe, P. W., & Muralidharan , K. 2016. 'Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications'. In Machin, S., Woessmann L., and Hanushek, E. A. (Eds.). *Handbook of the Economics of Education* Vol. 5. (pp. 653-743). Cambridge,MA: Elsevier.

Goldenberg, C. (2013). 'Unlocking the research on English learners'. *American Educator*, Summer 2013: 4-12.

Gove, A., and Wetterberg, A. (eds.) (2011). *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy*. Research Triangle Park, NC: RTI Press.

Guskey, T.R. (2002). 'Professional development and teacher change. Teach.' *Theory Pract*., 8 (3), 381–391.

Hanushek, E., and S. Rivkin (2010). 'Generalizations about Using Value-Added Measures of Teacher Quality'. *American Economic Review*, 100(2): 267–71.

Hargreaves, A. (2005). 'The emotions of teaching and educational change'. *In* Hargreaves, A. (*Ed*.), *Extending Educational Change: International Handbook of Educational Change*. Dordrecht: Springer. pp. 278–295.

Heckman, J.J. (2006). 'Skill Formation and the Economics of Investing in Disadvantaged Children'. *Science*, 312:1900-1902.

Hoadley, U. and Gallant, J. (2016) 'An Analysis of the Grade 3 Department of Basic Education workbooks as curriculum tools'. *South African Journal of Childhood Education,* 6(1): 1 – 28

Hoadley, U. (2012). 'What Do We Know about Teaching and Learning in South African Primary Schools?'. *Education as Change*, 16(2): 187–202.

Hoadley, U. (2016). 'A Review of the Research Literature on Teaching and Learning in the Foundation Phase in South Africa'. Research on Socioeconomic Policy (ReSEP) Working Papers: 05/16. Stellenbosch University. Available at: https://resep.sun.ac.za/a-review-of-the-research-literature-on-teaching-and-learning-in-the-foundation-phase-in-south-africa/ (accessed 15 April 2020).

Hoadley, U. (2018). *Pedagogy in Poverty: Lessons from Twenty Years of Curriculum Reform in South Africa.* Abingdon/New York: Routledge.Howie S, Combrink C, Roux K, Tshele M, Mokoena G, Palane N. (2017). '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.

Imbens, G.W., and D. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction* (Chapter 14)*.* Cambridge: Cambridge University Press.

Initial Teacher Education Research Project (ITERP). (2014). *An examination of aspects of initial teacher education curricula at five higher education institutions*. Johannesburg: JET Education Services.

Kennedy, M. (2016). 'How Does Professional Development Improve Teaching?' *Review of Educational Research*, 86(4): 945–80.

Kerwin, J.T., and R.L. Thornton (2019). 'Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures'. Available at: https://ssrn.com/abstract=3002723 (accessed 15 October 2019).

Kremer, M., Brannen, C., and Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, 340: 297-300.

Kotze, J., B. Fleisch, and S. Taylor (2019). 'Alternative Forms of Early Grade Instructional Coaching: Emerging Evidence from Field Experiments in South Africa'. *International Journal of Educational Development*, 66(1): 203–13.

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., D. Blazar, and D. Hogan (2018). 'The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence'. *Review of Educational Research*, 88(4): 547–88.

Lucas, A.M., P.J. McEwan, M. Ngware, and M. Oketch (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): 353–94.

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 (Accessed 5 October 2019).

Motala, S., and D. Carel (2019). 'Education Funding and Equity in South African Schools'. In N. Spaull and J.D. Jansen (eds), *South African Schooling: The Enigma of Inequality* (Chapter 4, pp. 67–86). Switzerland: Springer Nature.

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 . (accessed 1 December 2019).

NEEDU (2013). *NEEDU National Report 2012: The State of Literacy Teaching and Learning in the Foundation Phase*. Technical report. Pretoria: National Education and Evaluation Development Unit (NEEDU), Department of Basic Education.

Ouellette, G.P., and A. Haley (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. (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).

Piper, B., and M. Korda (2011). *Early Grade Reading Assessment (EGRA) Plus: Liberia: Program Evaluation Report*. Research Triangle Park, NC: RTI International.

Piper, B., D. Kwayumba, A. Oyanga, and E. Jepkemei (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., J. Destefano, E.M. Kinyanjui, and S. Ong'Ele (2018b). 'Scaling up successfully: Lessons from Kenya's Tusome National Literacy Program'. *Journal of Educational Change*, 19: 293– 321.

Piper, B., S.S. Zuilkowski, M. Dubeck, E. Jepkemei, and S.J. King (2018a). 'Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides'. *World Development*, 106: 324–36.

Piper, B., S.S. Zuilkowski, and S. Ong'Ele (2016). 'Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomised Control Trial in Kenya'. *Comparative Education Review*, 60(4): 776–807.

Piper, B., S.S. Zuilkowski, and A. Mugenda (2014). 'Improving Reading Outcomes in Kenya: First-Year Effects of the PRIMR Initiative'. *International Journal of Educational Development*, 37: 11–21.

Popova, A.; Evans, D.K.; Breeding, M. E.; and Arancibia, V. (2018). 'Teacher Professional Development around the World: The Gap between Evidence and Practice'. Policy Research Working Paper No. 8572. Washington, DC: World Bank.

Pretorius, E.J., and N. Spaull (2016). 'Exploring Relationships between Oral Reading Fluency and Reading Comprehension amongst English Second Language Readers in South Africa'. *Reading and Writing*, 29: 1449–71.

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

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

Reddy, V., C. Prinsloo, T. Netshitangani, R. Moletsane, and J. Van Rensberg (2010). *An Investigation into Education Leave in the South African Ordinary Public Schooling System*. Pretoria: Human Sciences Research Council.

SALDRU. (2019). *Funda Wande Baseline and Midline Evaluations*. [Data file]. Cape Town: SALDRU, University of Cape Town.

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

Shalem, Y. and De Clerq, F. (2019) 'Teacher Development and Inequality in Schools: Do We Now Have a Theory of Change'. *In* Spaull, N. and Jansen, J.D. (*eds*.) *South African Schooling: The Enigma of Inequality*. Springer. Chapter 13.

Snilstveit, B., J. Stevenson, R. Menon, D. Philips, E. Gallagher, M. Geleen, H. Jobse, T. Schmidt, and E. Jimenez (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 N. Kucirkova, C. Snow, V. Grøver, and C. McBride-Chang (eds), *The Routledge International Handbook of Early Literacy Education: A Contemporary Guide to Literacy Teaching and Interventions in a Global Context*. New York, NY: Routledge.

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

Spaull, N. (2015). 'Accountability and capacity in South African education'. *Education as Change,* 19(3), 113–142.

Spaull, N. (2011). *Primary School Performance in Botswana, Mozambique, Namibia and South Africa: A Comparative Analysis of SACMEQ III.* South African Consortium for Monitoring Educational Quality.

Spaull, N. and U. Hoadley (2017). 'Getting reading right: Building firm foundations' In Jamieson, L., Berry, L. and Lake, L. (Eds.) *South African Child Gauge 2017*. (pp. 77-83). Cape Town, Children's Institute: University of Cape Town.

Spaull, N., and J. Kotze (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., and E. Pretorius (2019). 'Still Falling at the First Hurdle: Early Grade Reading in South Africa'. In N. Spaull and J.D. Jansen (eds), *South African Schooling: The Enigma of Inequality* (Chapter 8, pp. 147–68). Switzerland: Springer Nature.

Spaull, N., E. Pretorius, and N. Mohohlwane (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. (2013). Poverty and privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, 33(5): 436–447.

Spear-Swerling L. (2006). 'Children's reading comprehension and oral reading fluency in easy text'. *Reading and Writing: An Interdisciplinary Journal,* 19: 199–220.

Taylor, N. 2019. Inequalities in Teacher Knowledge in South Africa. *In* Spaull, N. and Jansen, J.D. (*eds*.) South African Schooling: The Enigma of Inequality. Springer. Chapter 14.

Taylor, N. 2014. *N*EEDU National Report 2013: Teaching and Learning in Rural Primary Schools. Pretoria: government Printer

Taylor, S. (2011). 'Uncovering indicators of effective school management in South Africa using the National School Effectiveness Study'. Stellenbosch Economic Working Papers 10/11. Stellenbosch University. Available at: https://resep.sun.ac.za/wp-content/uploads/2017/10/wp-10-2011.pdf (accessed 5 October 2020).Taylor, N., and S. Taylor (2013). 'Teacher Knowledge and Professional Habitus'. In N. Taylor, S. Van den Berg, and T. Mabogoane (eds), *Creating Effective Schools: Report of South Africa's National Schools Effectiveness Study* (pp. 31–60). Cape Town: Pearson.

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., Cilliers, J., Prinsloo, C., Fleisch, B., Van der Berg, S. and Reddy, V. (2016). 'The Early Grade Reading Study: Midline Report'. Department of Basic Education: Pretoria. [Midline Report].

Taylor, S., and M. Von Fintel (2016). 'Estimating the Impact of Language of Instruction in South African Primary Schools: A Fixed Effects Approach'. *Economics of Education Review*, 50: 75–89.

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

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., N. Spaull, G. Wills, M. Gustafsson, and J. Kotze (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) initiative. Pretoria: PSPPD.

Van der Berg, S. (2008). 'How Effective Are Poor Schools? Poverty and Educational Outcomes in South Africa'. *Studies in Educational Evaluation*, 34(3):145-154.

Venkat, H. and Spaull, N. (2015) 'What do we know about primary teachers' mathematical content knowledge in South Africa? An analysis of SACMEQ 2007'. *International Journal of Educational Development*, 41:121-130.

What Works Clearinghouse. (2020). *Standards Handbook*. Version 4.1. Washington, D.C.: Institute of Education Sciences. Available at: https://ies.ed.gov/ncee/wwc/Handbooks (accessed 15 October 2020).

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*. Washington, DC: Systematic Country Diagnostic.

Zuilkowski, S.S., and B. Piper (2017). *Instructional Coaching in Kenya: Supporting Teachers to Improve Literacy Outcomes*. In M. Akiba and G.K. LeTendre (eds), *International Handbook of Teacher Quality and Policy* (pp. 505–16). New York, NY: Routledge.

9 APPENDIX

9.1 APPENDIX TABLES AND FIGURES

TABLE A1: BALANCE IN GRADE 1 EGRA SCORES

	Treatment			Control			p-value
	Mean	s.d.	N	Mean	s.d.	N	
Receptive Listening Comprehension	9,55	0,85	301	9,35	1,19	294	0,14
Productive Listening Comprehension	2,73	1,54	301	2,58	1,53	294	0,36
Expressive Vocabulary	10,35	3,91	301	10,95	4,18	294	0,22
Letter Sounds per minute	5,00	9,07	301	6,01	9,60	294	0,31
Digraphs and Trigraphs per minute	0,15	1,21	301	0,24	1,37	294	0,55
Phonemic Awareness	1,41	1,85	300	1,61	1,80	294	0,30
Word Choice	1,68	1,71	301	1,90	1,58	293	0,19
Rapid Automatized Naming	12,30	3,92	301	12,50	3,89	294	0,61
Write your name	4,49	1,05	298	4,59	0,82	294	0,37
Copy a word	4,33	1,41	298	4,32	1,37	294	0,92
Write letters	1,36	1,46	298	1,73	1,51	294	0,08

TABLE A2: BALANCE IN GRADE 2 EGRA SCORES

	Treatment			Control			p-value
	Mean	s.d.	N	Mean	s.d.	N	
Receptive Listening Comprehension	9,71	0,61	302	9,61	0,83	290	0,19
Productive Listening Comprehension	3,47	1,36	302	3,61	1,33	290	0,31
Expressive Vocabulary	12,15	3,45	302	12,36	3,84	290	0,63
Letter Sounds per minute	28,42	20,27	302	28,48	19,53	290	0,98
Digraphs and Trigraphs per minute	8,88	11,44	302	8,10	11,25	290	0,62
Phonemic Awareness	4,21	2,31	302	4,32	2,39	290	0,72
CVCV Words per minute	10,17	12,29	302	9,41	11,80	290	0,65
Familiar Words per minute	7,11	8,93	302	6,54	8,59	290	0,64
Oral Reading Fluency	7,53	9,71	302	7,33	9,07	290	0,88
Reading Comprehension	4,13	4,82	302	4,27	4,59	290	0,83
Vocabulary	3,08	2,40	301	3,19	2,25	290	0,74
Sentence Comprehension	4,44	4,46	302	4,52	4,34	290	0,90
Write letters	3,93	1,42	292	4,01	1,26	290	0,65
Write words	13,79	7,11	292	14,47	6,53	290	0,52

TABLE A3: BALANCE IN LEARNER CHARACTERISTICS AND HOME ASSETS (BOTH GRADES)

	Treatment			Control			p-value
	Mean	s.d.	N	Mean	s.d.	N	
Female	0,51	0,50	603	0,49	0,50	584	0,58
Age in months	84,16	10,66	603	83,37	10,21	584	0,31
Wearing spectacles	0,02	0,14	603	0,02	0,14	584	0,92
Height (cm)	119,20	6,82	603	118,93	6,64	584	0,59
Absent >= 1-day past week	0,27	0,44	596	0,29	0,46	577	0,42
Teacher absent >= 1-day past week	0,23	0,42	563	0,24	0,43	545	0,70
Library in the school	0,36	0,48	489	0,37	0,48	481	0,89
Readers or storybooks in classroom	0,87	0,34	597	0,92	0,27	580	0,04
Lives with mother	0,89	0,31	603	0,91	0,29	584	0,44
Lives with father	0,62	0,49	603	0,65	0,48	584	0,28
Lives with both parents	0,59	0,49	603	0,62	0,48	584	0,33
Lives with neither parents	0,08	0,27	603	0,06	0,24	584	0,26
Books (not schoolbooks) at home	0,32	0,47	602	0,38	0,48	583	0,12
Radio	0,70	0,46	602	0,72	0,45	584	0,66
Mobile	0,99	0,10	602	0,99	0,11	584	0,82
Electricity	0,97	0,17	602	0,96	0,19	583	0,61
Television	0,95	0,23	602	0,94	0,24	583	0,70
Computer	0,30	0,46	603	0,33	0,47	578	0,34
Fridge	0,91	0,28	602	0,94	0,23	583	0,12
Toilet	0,59	0,49	601	0,61	0,49	583	0,54
Bicycle	0,45	0,50	518	0,40	0,49	534	0,30
Vehicle	0,50	0,50	601	0,53	0,50	584	0,32

TABLE A4: 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 A5: SUB-TASK MIDLINE RAW SCORE DISTRIBUTIONS, BY GRADE AND TREATMENT STATUS

	Control									Treatment								
Task	N	% zero	Mean	s.d.	Mean (excl. 0)	Percentiles 25th	50th	75th	max	N	% zero	Mean	s.d.	Mean (excl. 0)	Percentiles 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

TABLE A6: ORDERED LOGIT RESULTS FOR TREATMENT EFFECT ON TEACHER ABILITY TO IDENTIFY MOST/LEAST PROFICIENT READERS BASED ON OBJECTIVE CLASS RANK

	Identified "Most proficient"	Identified "Least proficient"
Treatment	-0.665**	-0.423
T=0 p-value	0.050	0.293
Observations	103	106

9.2 ANNEX I: SUMMARY OF THE EGRS INTERVENTIONS AND RESULTS COMPARISON

9.2.1 WHAT WE LEARNT FROM EGRS I AND II RESULTS

The two government-run Early Read Grade Reading studies are of particular relevance, as an important domestic comparative intervention from which insights were used to design the Funda Wande intervention. The first round of the Early Grade Reading Study (EGRS I) was evaluated using a combination of an RCT, detailed classroom observation, and case studies conducted between 2015-2017 in 230 quintile 1-3, no-fee schools in the North West province of South Africa (Cilliers et al., 2019a; Taylor et al., 2017). Both of the key treatment arms discussed in Cilliers et al. (2019a) provide teachers with lesson plans aligned to the national curriculum, as well as reading materials (booklets, flash cards and posters). The major difference between these two treatment arms[90] is that i) in one group training takes place at a centralised workshop twice a year (the status quo teacher professional development model), and ii) in the other training takes place in small clusters with ongoing support for teachers by coaches who observe lessons in classrooms, provide feedback, and demonstrate effective teaching techniques in-situ.

The authors estimate 0.12 and 0.24 s.d. increases in a composite measure of reading proficiency in the Training and Coaching groups, respectively. The coaching intervention also brought about the larger relative shift in reading comprehension (as arguably the most important goal of literacy interventions), also positive impacting on all the literacy sub-outcomes measured. As far as effects sizes for educational interventions are concerned, the impact of the coaching intervention was relatively large and effective in shifting learning outcomes[91].

Other positive findings include the fact that the programme seems to allows boys to narrow the significant gap in reading proficiency vis-à-vis girls, and that relatively large classes (38-45 learners) saw the largest impacts. Impacts were also concentrated in urban schools More discouraging is the finding that the weakest learners at baseline did not benefit from the programme, with impacts from the programme largest for learners in the middle-to-top of the baseline reading proficiency distribution.

Turning to the potential mechanisms at play, a combination of teacher surveys and classroom observations indicate that teachers in the intervention arm felt more supported, their classrooms had more access to print resources, and treated teachers adhered more closely to the curriculum prescribed routine. Overall, the interventions shifted teacher's general instructional practice. The most important mechanism, as carefully argued by Cilliers et al. (2019a), is the use of more technically demanding teaching techniques. Coaching arm teachers, in particular, were more likely to use techniques like group guided reading. In turn, their learners are more likely to receive individualised attention when reading, and also spend more time actually using the reading materials provided (Cilliers et al., 2019a).

Noteworthy is that that even though the coaching intervention is more expensive in absolute terms, it is more cost effective in terms of learning gains per USD spent when evaluated over two years (with 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,

[90] The average duration of exposure to the programmes were roughly equivalent (Cilliers et al., 2019:3).
[91] See Kraft (2019) for an updated schema of the literature on effect size benchmarks for education programmes.

however, it is unclear whether the coaching programme is still more cost-effective (Cilliers et al., 2019b). Nevertheless, the coaching intervention did perform relatively better on certain key impact measures w.r.t. 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 learner 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 learners' 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 learner learning outcomes, requires some form of ongoing in-classroom support, monitoring and feedback for teachers, like that offered by the coaches in the intervention

9.2.2 In depth comparison of Funda Wande and EGRS midline results

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). In the second year when the evaluation cohort was 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 case in the EGRS 1 evaluation after either one or two years of programme implementation. These results suggest that a) it is learners' 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 learners 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 learners 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.

9.3 ANNEX II: FUNDA WANDE PROGRAMME INPUTS

FIGURE B1: LTSM BOX EXHIBIT

FIGURE B2: GRADE SPECIFIC LESSON PLANS

FIGURE B3: LESSON PLAN INTRODUCTION SECTION

Big Idea 3 : Using the Funda Wande lesson plans

Using the Funda Wande lesson plans

What is the purpose of the Funda Wande lesson plans?

The purpose of these lesson plans is to give teachers strategies to teach large classes of varied learning abilities, without leaving anyone behind.

Video #3: Introduction to the Lesson Plans. (3:

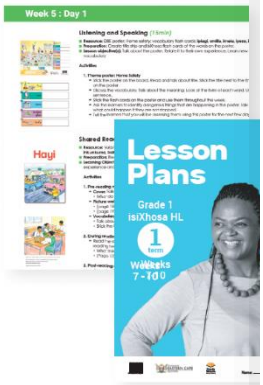
What is in a lesson plan?

Each lesson has the following:

1. Resources for the lesson
2. Preparation for the lesson
3. Objectives of the lesson
4. Step-by-step guide
5. Formal Assessment Tasks (FATs) per learning area per term

What are the resources provided with the lesson plans?

1. **Resources for the lesson**
   - DBE posters
   - Flash cards
   - Read Aloud - Storybook
2. **Shared Reading**
   - Vula Bula Big Books
   - Flash cards
   - Story elements poster
   - Comprehension strategies cards
   - Vocabulary charts
3. **Phonics**
   - Vula Bula alphabet frieze
   - Vula Bula alphabet words
   - Vula Bula phonics frieze
   - Vula Bula phonics words
   - Phonics chart
4. **Handwriting**
   - Handwriting booklet (Grade 1)
5. **Writing**
   - Writing frame poster
6. **Group Guided Reading**
   - Baseline Assessment booklet
   - Group Guided Reading booklet



What is special about the Funda Wande lesson plans?

1. In **Listening and Speaking**, we teach the following oral skills:
   - Debating
   - Interviews
   - Retelling
   - Reciting.

We also provide opportunities for each learner to speak through a guided activity that involves pair talk and using sticks to give every learner a turn to talk.

Video #4: Components of Lesson Plans. (5:43 min)

2. In **Shared Reading**, we teach the following:
   - Grammar
   - Punctuation
   - Comprehension strategies.
3. In **Phonics**, we provide opportunities for a learner to identify and manipulate sounds.
4. In **Writing**, learners are given sufficient support to master the skill of writing through:
   - Shared Writing **(Modelled)**
   - Paired Writing **(Peer support)**
   - Independent Writing.
5. In **Handwriting**, we assist the teacher by providing the handwriting booklet for learners to have more practice.
6. In **Group Guided Reading**, we have compiled a Group Guided Reading booklet that explicitly shows:
   - How to conduct Baseline Assessment in preperation for GGR groups.
   - How to set up groups and a timetable for GGR.
   - How to establish GGR routines.
   - How to conduct GGR lessons using Vula Bula anthologies.
   - Lastly, how to use the rubric to assess reading in GGR.

What is the teacher's role?

The role of the teacher is to:

- Familiarise herself with the structure of the lesson plans.
- Read the stories in advance and prepare open-ended questions.
- Cut out flash cards provided for each lesson.
- Try to stick to the lesson plan as much as possible.
- Enjoy your lessons and have fun with your learners!

FIGURE B4: LESSON PLAN DAILY OUTLINE EXAMPLE

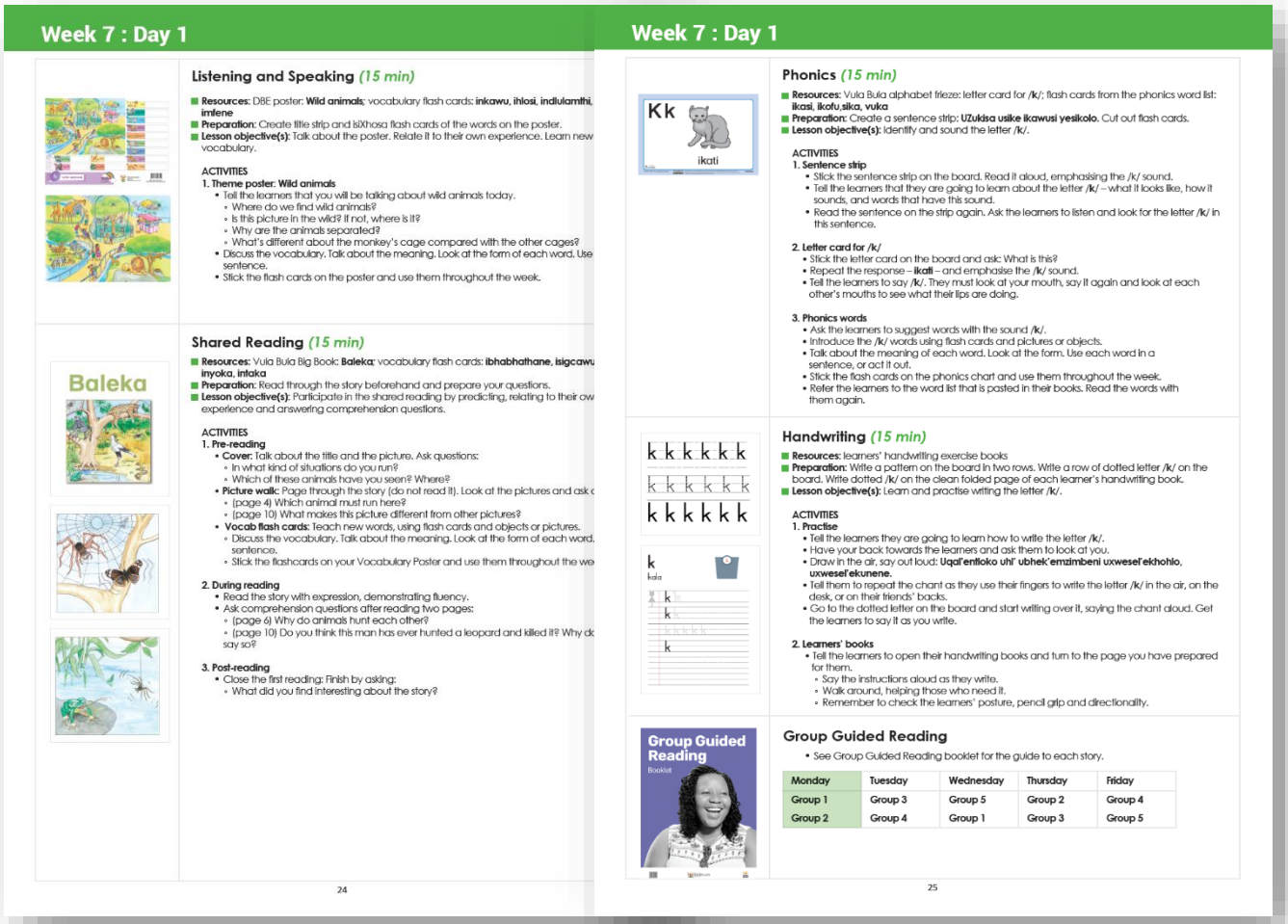


FIGURE B5: ONLINE TEACHER RESOURCES (YOUTUBE TRAINING VIDEOS)

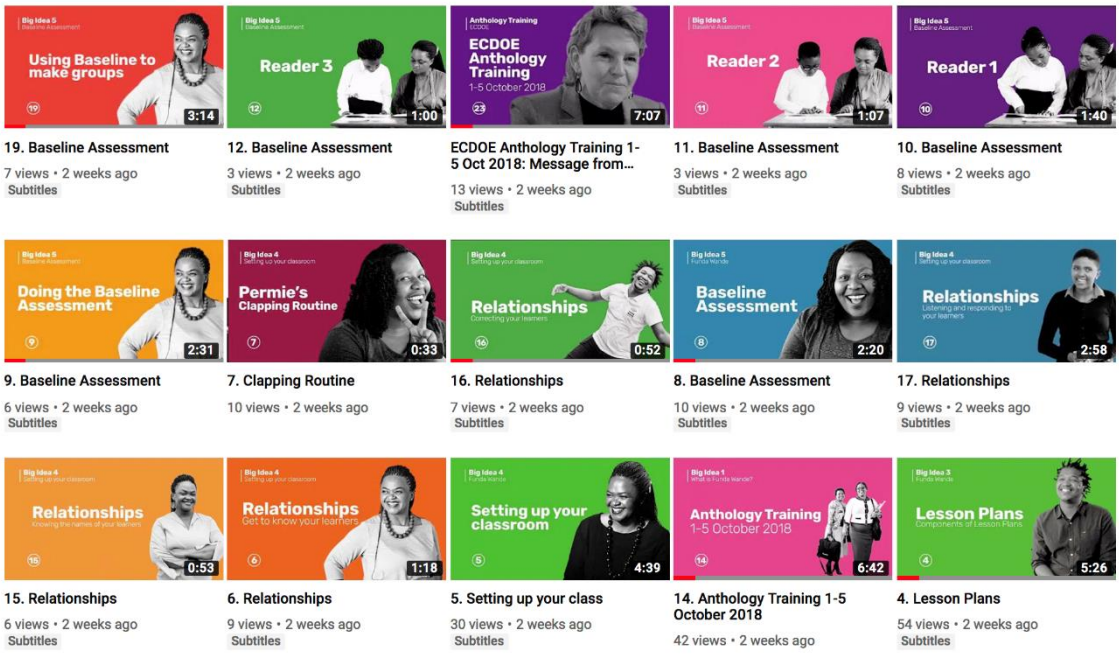


FIGURE B7: BASELINE ASSESSMENT-, HANDWRITING-, AND GROUP GUIDED READING BOOKLETS

FIGURE B8: GROUP GUIDED READING BOOKLET CONTENT EXAMPLE

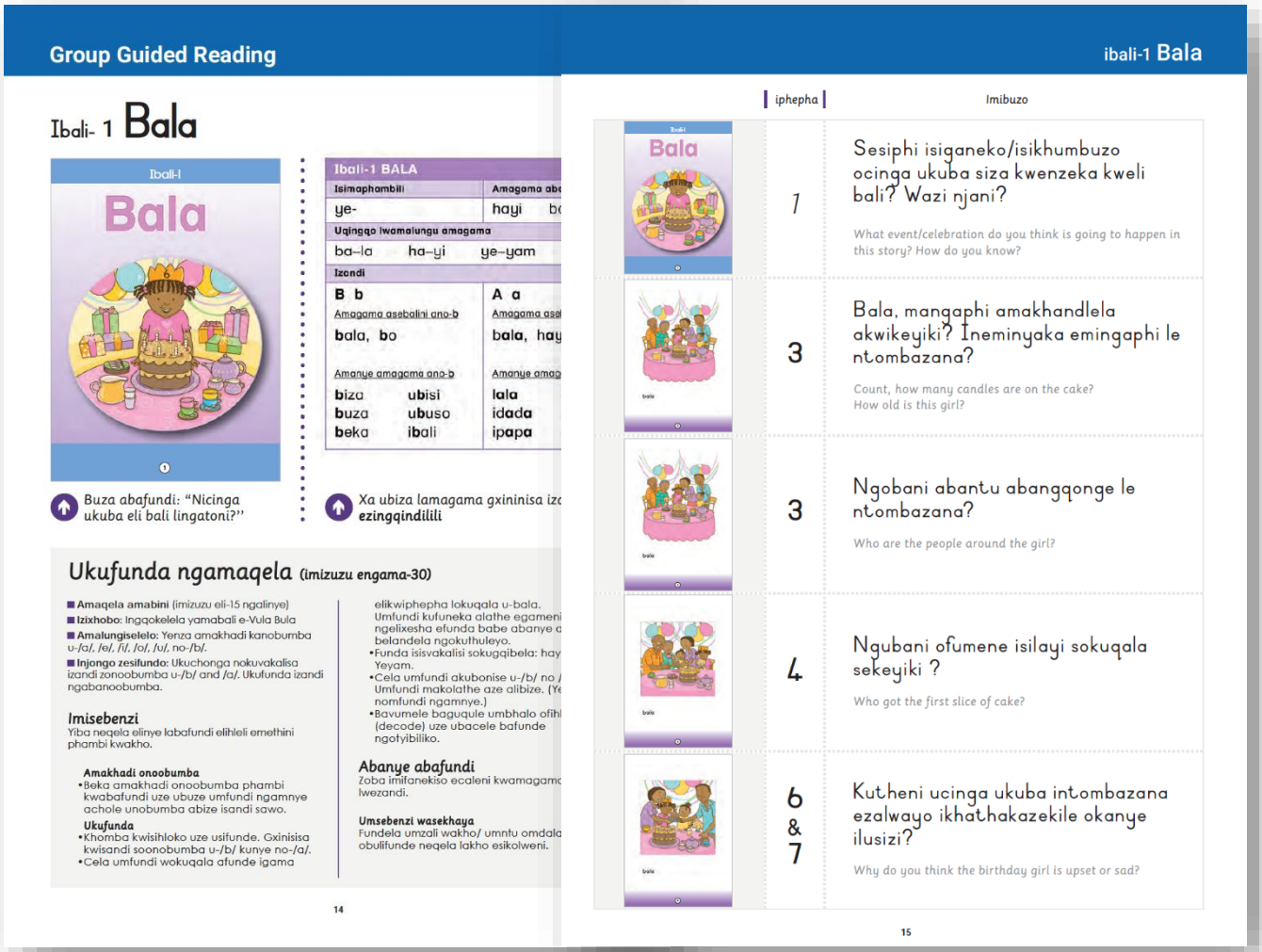
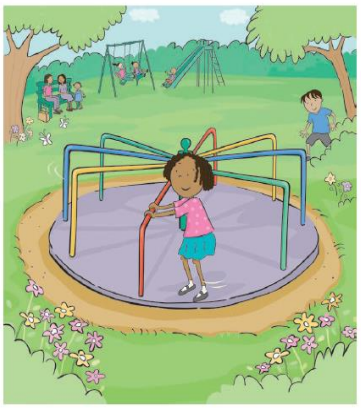


FIGURE B9: VULA BULA GRADED READERS

Hayi, Velile! Suku...



Ujikeleza uqalile ukujika-jika.
Uyajika-jika.

Nalo uza kujikeleza.

Ubona umfanekiso wenyoka encwadini yolwazi lwezilwanyana. Upopola amazinyo enyoka ...

Asililo izinyo lenyoka. Amazinyo enyoka awafani neli zinyo.

FIGURE B10: TRAINING ON USE OF VULA BULA GRADED READER

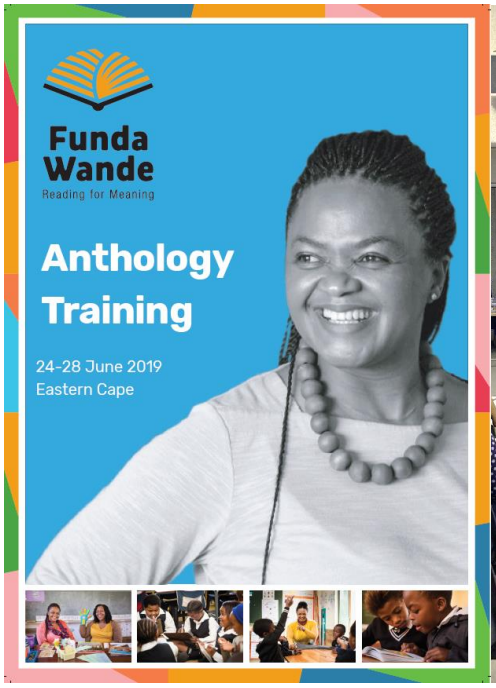
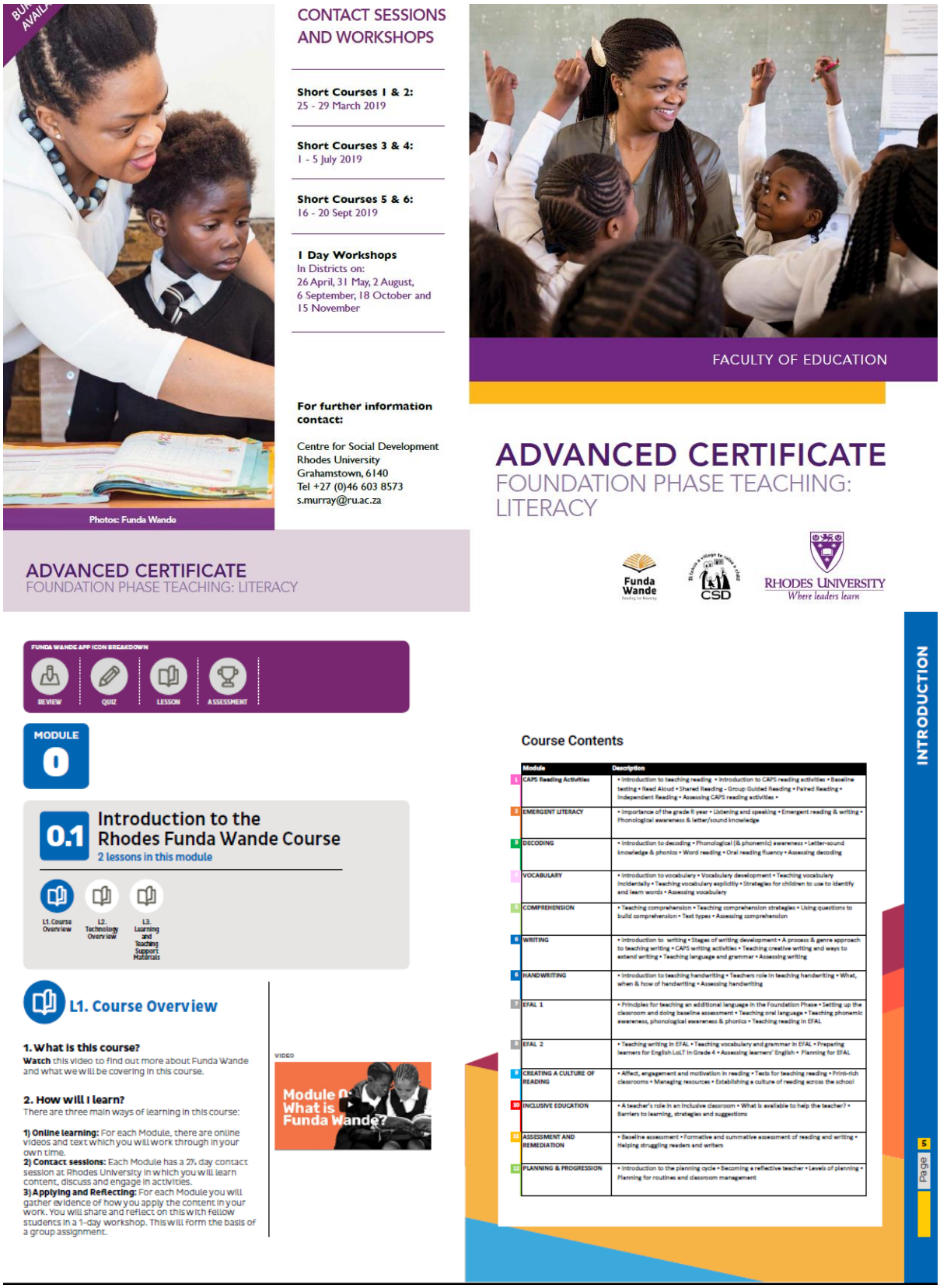

Funda
Wande
Reading for Meaning
Anthology
Training
24-28 June 2019
Eastern Cape

FIGURE B11: ADVANCED CERTIFICATE IN TEACHING FOUNDATION PHASE LITERACY FROM RHODES UNIVERSITY



9.4 ANNEX III: SCHOOL SELECTION AND UNDERLYING POPULATION

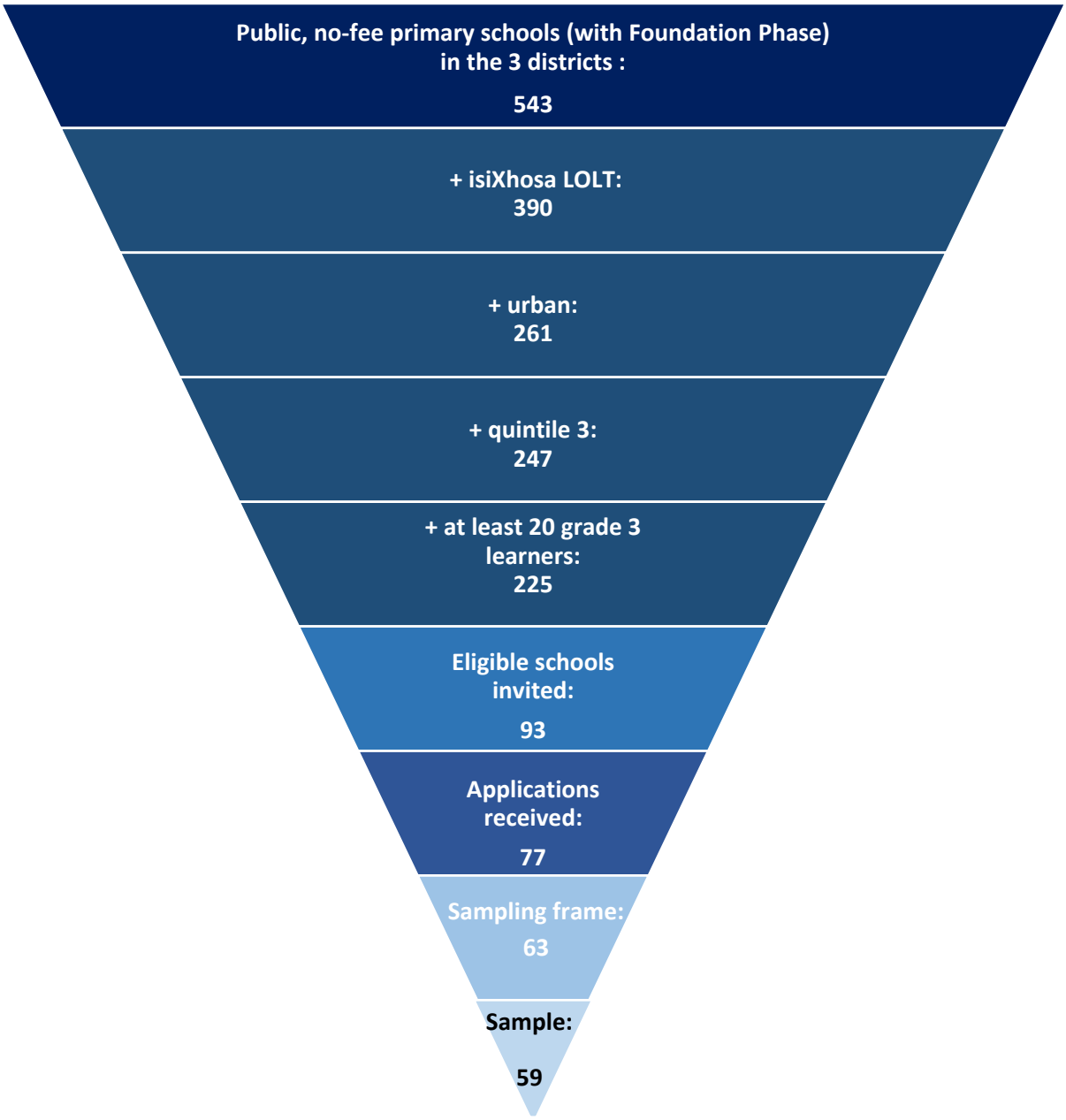
Based on the primary inclusion criteria, Funda Wande received a list of eligible schools from the three respective district managers. The list of 93 schools[92] were based on the explicit criteria that schools should be no-fee, public primary or -combined schools (i.e. have Grade 1-3 learners), with 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 C1 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[93]. Of the 543 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.

[93] 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).

Figure C1: School selection process from plausible populations of schools in the 3 districts



With respect to creating the three strata within which randomisation took place, 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, with two of these schools randomly selected as possible replacement schools - randomly assigning one to treatment and the other to control. Within each group (or strata) of 20 schools, half of the schools were randomly assigned to receive the Funda Wande program, with the other half serving as control schools.

Furthermore, post randomization it was discovered that the LOLT of two control schools was not isiXhosa throughout the Foundation Phase. These schools were subsequently dropped and one school from the replacement group was added to the treatment group.

9.5 ANNEX IV: LEARNER ASSESSMENT DETAILS AND RATIONALE

Many of the extended Early Grade Reading Assessment (EGRA) task used in the Funda Wande evaluation instrument built on minor adaptations made by NORC, at the University of Chicago, for the Story Powered Schools Impact Evaluation[94]. The instrument development also benefited hugely from the input of the Story Powered Schools and Funda Wande evaluation field teams, particularly on translations and appropriate language.

The range of literacy and pre-literacy assessments conducted at baseline were generally used again for the midline learner evaluations (see table D1). 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 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, baseline writing tasks were also excluded at midline[95].

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 newly introduced passage provides a second measure of reading fluency and subsequent reading comprehension assessment for Grade 2 learners. Having two different texts on which reading fluency and -comprehension are assessed allows 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 recalling the texts).

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,

[94] 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).

[95] Writing tasks are likely to be included in some of the subsequent rounds of data collection, as learners become more proficient readers and the assessment of higher order abilities (like creative writing) become increasingly important to differentiate among the better performing learners.

before they were asked comprehension questions based on the respective passages immediately thereafter.

TABLE D1. 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 learner	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 learner		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	

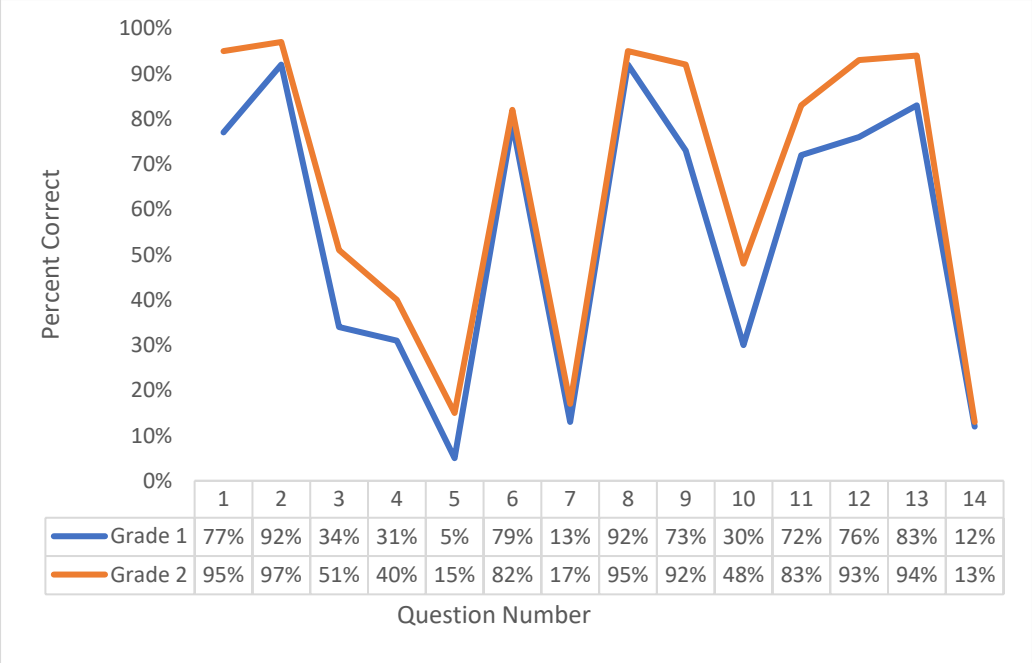
In light of the challenge presented by low reading fluency levels, another task was included to assess learners' reading comprehension whilst relying significantly less on their fluency levels. More precisely, the sentence comprehension subtask was untimed and consisted of 20 short sentences (typically 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.

In contrast, three separate reading comprehension tasks were employed, 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).

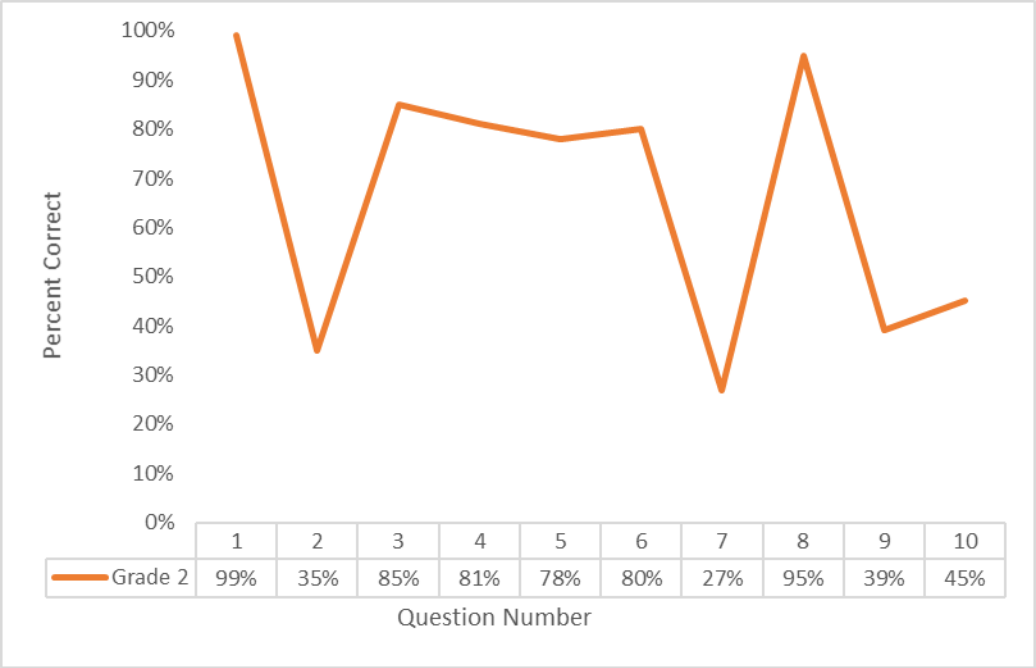
Figures D1 and D2 indicate the percentage of learners who could correctly answer each question of the two paragraph reading comprehension tasks, but only for those learners who finished reading the whole passage in three minutes. There is a large variability in learners' ability to answer the questions 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 D1: 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 D2: HOW LEARNERS WHO ATTEMPTED ALL QUESTIONS FARED BY QUESTION (READING COMPREHENSION II, GRADE 2 ONLY)

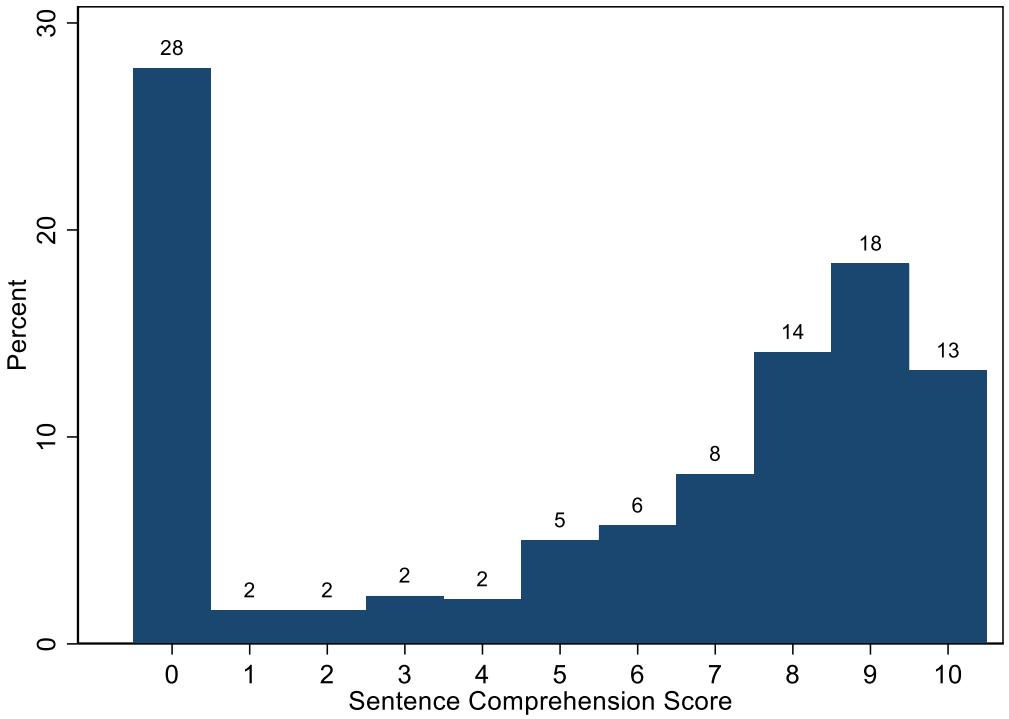


	1	2	3	4	5	6	7	8	9	10
Grade 2	99%	35%	85%	81%	78%	80%	27%	95%	39%	45%

Figure D3 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 D3: HISTOGRAM OF SENTENCE COMPREHENSION SCORES (GRADE 2 ONLY)



9.6 ANNEX V: INTERMEDIATE OUTCOMES – TEACHER FORMATIVE ASSESSMENT (FURTHER ANALYSIS)

To assess the extent to which teachers in control and intervention schools were attuned to the actual reading levels of the learners in their classrooms (an important mechanism by which the Funda Wande intervention aims to improve reading outcomes), teachers were presented with three reading passages at the Grade 1, Grade 2 and Grade 3 levels, respectively. These passages drew from graded readers provided in equal measure to teachers in both control and treatment schools at the start of the school year. Based on the passage presented to them, teachers were asked how many learners in their class would be able to read the 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 E1). 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, whilst one in two of their learners would manage to read the Grade 3 text on their own.

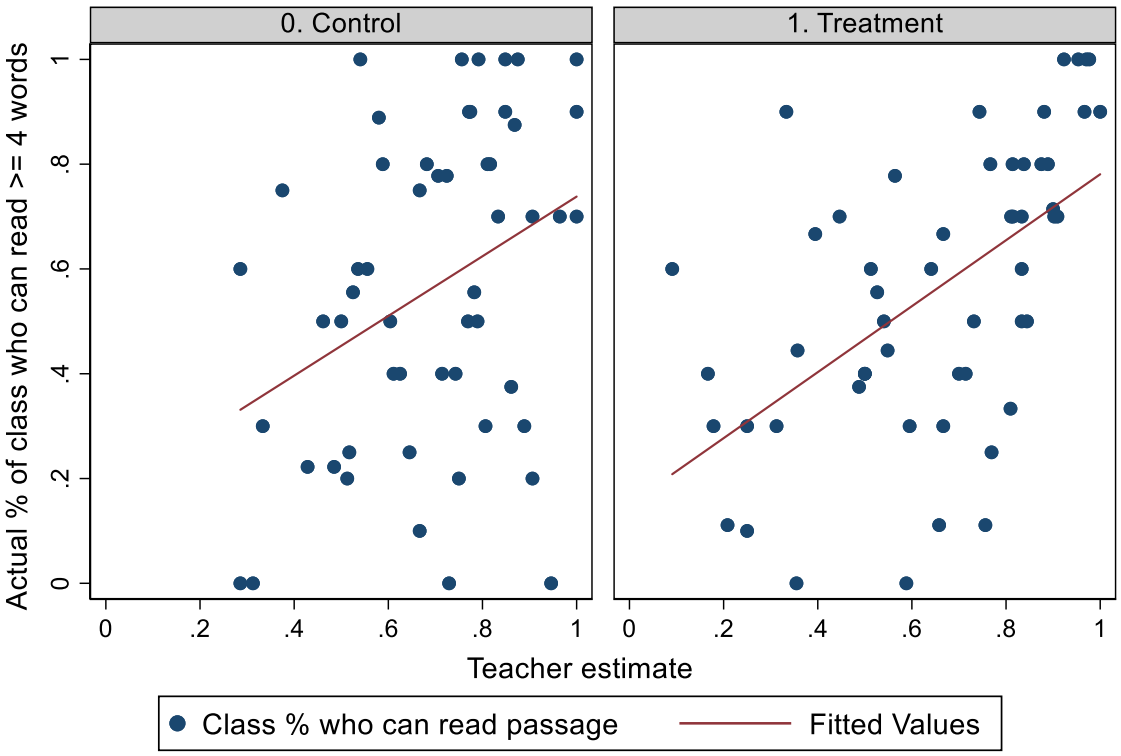
TABLE E1: PERCENTAGE OF LEARNERS IN CLASS ABLE TO READ TEXT AT THIS LEVEL

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

Based on the data for the 10 randomly selected learners per class, one can compare how 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 at least four words correctly, whilst the Grade 2 passage consists of 18 words. This is used to construct two indicators of the share of the 10 assessed learners in each teachers' class that can read at least 4 and 18 words, respectively.

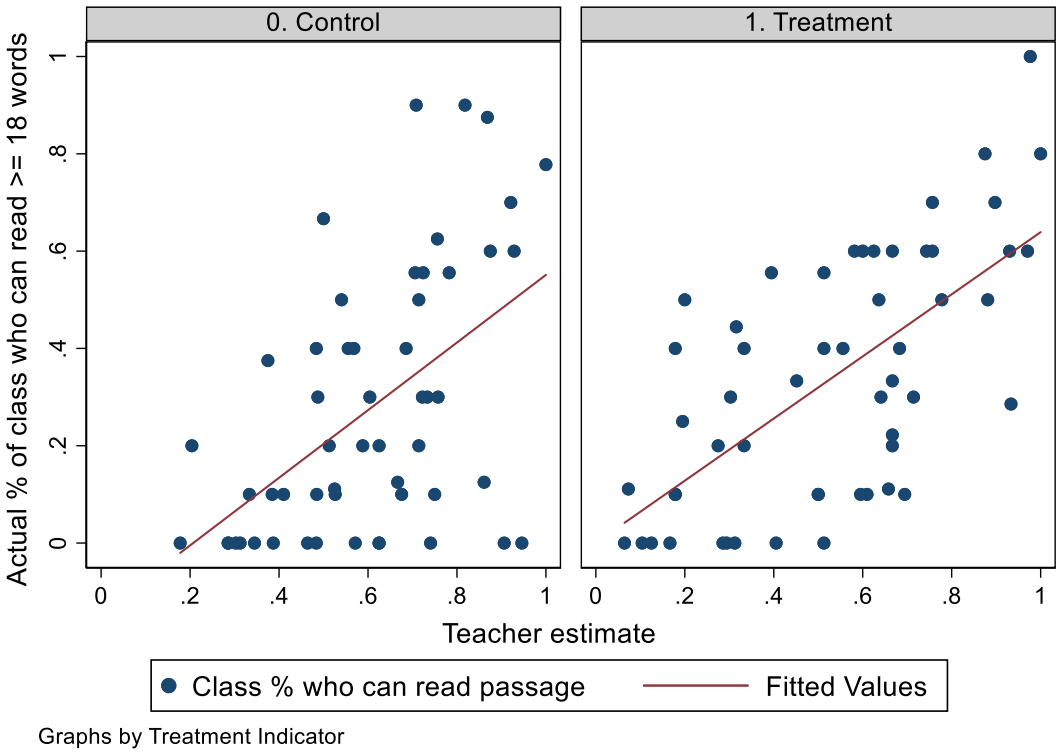
Figure E1 below graphically presents how well teachers' estimates of the share of the class who can read the grade 1 level passage (on the horizontal axis) compared to the share of learners assessed who achieved minimum fluency levels required to read the passage (on the vertical axis), for classrooms 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 assessed 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 E2.

FIGURE E1: PERCENT OF LEARNER WHO CAN READ >=4 WORDS PER MINUTE OVER TEACHER REPORT ON PERCENT WHO CAN READ GRADE 1 PASSAGE, BY TREATMENT STATUS



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

FIGURE E2: PERCENT OF LEARNER WHO CAN READ >=18 WORDS PER MINUTE OVER TEACHER REPORT ON PERCENT WHO CAN READ GRADE 2 PASSAGE, BY TREATMENT STATUS



Despite intervention school teachers being the better predictors, the predictions in both groups were fairly inaccurate. This can be gleaned from the low R-squared values, but also simply from inspecting Figures E1 and E2. 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 E2 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).