THE USE OF EARLY-GRADE READING BENCHMARKS TO IMPROVE THE EFFICACY OF ASSESSMENT IN AFRICAN LANGUAGES

Asanda Gontse Lobelo

A dissertation submitted to the Faculty of Commerce in partial fulfilment of the requirements for the degree of Master of Commerce specialising in Applied Economics

December 2024

School of Economics
Faculty of Commerce
University of Cape Town

Supervisor: Professor Cally Ardington

Abstract

South Africa is facing a ‘reading crisis’ where the majority of learners are unable to read for meaning or with understanding by the end of Grade 4. In recognizing this crisis, there have been considerable sector-wide efforts to improve foundational literacy outcomes, including the establishment of reading benchmarks in African languages. These benchmarks are a measure of grade-level reading proficiency that can be used to monitor progress at the national, provincial, district and school level. One way the benchmarks could be used productively at the classroom level, is enabling teachers to interpret assessment results into learning levels, based on progress towards meeting the benchmark.

Formative assessment is crucial for effective teaching and implementing the curriculum. Furthermore, differentiated instruction programs have been gaining traction for their demonstrated potential to improve learning outcomes in contexts where within-grade heterogeneity is high with many learners not keeping pace with the curriculum. A key assumption of these programs is that educators know the learning levels of their learners. The evidence around teacher formative assessment practices and the efficacy thereof suggests that this assumption is not likely to hold in the South African context. To create a basis for differentiated instruction programs to improve literacy outcomes, teacher knowledge of the learning levels of their learners needs to be evaluated. Secondly, the processes through which teachers gain that knowledge, the formative assessment process, needs to be strengthened.

The overarching purpose of this study is to generate insights into how the newly established benchmarks could be productively used in South African classrooms. We do this by examining current teacher knowledge of learning levels through their existing formative assessment practices and evaluate the effectiveness of a benchmarks-orientated intervention in improving that knowledge. Using longitudinal data from this pilot study across 39 schools, this paper estimates the intent-to-treat (ITT) effect, local average treatment effect and average treatment effect on the treated (ATT) of the intervention. The results show that teachers tend to overestimate the performance of their learners across the achievement distribution and the size of the misestimation is relatively large. Take-up rates of the intervention were fairly low and variable, but we find evidence that the intervention improved knowledge of relative reading proficiency of intervention-trained teachers relative to their untrained counterparts. This is even more so for trained teachers for whom we are able to confirm use of the intervention materials. These findings underscore the importance of enhancing the resources, training, and support provided to teachers for the formative assessment process. Furthermore, the inclusion of the reading benchmarks could be useful in augmenting such support.

UNIVERSITY OF CAPE TOWN
FACULTY OF COMMERCE
Igniting Knowledge and Opportunity

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	LBLASA001
Student name	Asanda Lobelo
Signature of Student	Signed by candidate
Date:	05/12/2024

Acknowledgements

I would like to thank Professor Cally Ardington for her generous mentorship, support and supervision. Through her work, I have been inspired to contribute to knowledge production in the South African education sector and the efforts to address fundamental challenges in improving foundational literacy and numeracy. I am incredibly grateful to have worked with someone with such a depth and breadth of knowledge and whose passion inspires me to challenge myself and be curious. Her kindness, patience and compassion in supporting me to navigate various challenges, both personal and professional, has been invaluable in the process of writing this dissertation.

I would like to acknowledge scholarship funding from the National Research Foundation, SALDRU and Funda Wande through the NextGen Fellowship. I also acknowledge project funding from the Zenex Foundation and J-PAL Africa as well as the partnership with the Department of Basic Education, without which this research would not have been possible.

Finally, I dedicate this dissertation to my parents, Mandisa and Othusitse. I am because they are.

Table of Contents

1. Introduction

Despite having almost universal primary school enrolment, school quality and learner performance remain low in South Africa, particularly in schools that serve disadvantaged communities (van der Berg et al., 2011). Improving school quality in low-income contexts is an important policy question as access to quality education is key for poverty alleviation and social mobility. In recognising the need to improve the quality of education in South Africa, President Cyril Ramaphosa issued a directive that all children should be able to read for meaning by the age of 10 (Ramaphosa, 2019). The challenge for policymakers at the Department of Basic Education (DBE) has been to define what reading for meaning tangibly means in terms of quantitative measures of proficiency in foundational reading skills across the 11 official languages. As such, DBE, in collaboration with various stakeholders, has been leading efforts to establish reading benchmarks in African languages.

Teachers arguably bear the biggest responsibility in meeting this new directive. They play an important role in student performance because they facilitate literacy development on a daily basis. Teaching literacy in the absence of well-defined targets and milestones means that teachers potentially do not know where their learners are and where they should be in their fluency development. Benchmarks have the potential to improve teachers' understanding of learners' reading proficiency so that they are better able to support learning through tailoring their instruction. The next phase of the benchmarking efforts is to consider how to effectively disseminate these newly established benchmarks to teachers for practical use in their classrooms. These benchmarks serve as a reliable yardstick against which teachers can gauge their learners' progress. When a student's reading proficiency falls below the benchmark, it serves as a clear indicator of their need for additional support. One particularly valuable application of benchmarks in the classroom is within the formative assessment process. Formative assessment is where teachers use assessment results to identify learning needs and make adjustments to their instructional approach to meet those needs (William & Thompson, 2007). Thus, benchmarks become an important lens through which teachers can quantitatively interpret assessment results and translate them into actionable insights regarding their students' learning needs.

In order for benchmarks to be productively used in this way, teachers need to have existing, well-defined formative assessment processes. Although teachers are required to conduct formative assessments as part of the National Assessment Protocol (Department of Basic Education, 2012), they are not provided with a framework to do this. Recent developments, like the national rollout of the Early Grade Reading Assessment (EGRA) as an in-classroom

diagnostic tool, have been important in standardising assessments however the roll out is yet to reach all schools. Furthermore, the EGRA roll-out assumes that teachers are able to interpret EGRA scores in the context of the curriculum and differentiate teaching approaches in response. While there is a growing body of research on in-service teacher development to improve pedagogic practice in South Africa, less is known about the formative assessment practices of teachers. Existing research suggests that teachers do not have a good grasp of formative assessment (Mkhwanazi et al., 2014; Kanjee & Mthembu, 2015) and do not know the proficiency levels of learners in their class (Hadjidemetriou & Williams, 2001; Maunganidze et al., 2008; Kilday et al., 2012; Djaker et al., 2022). Without knowing the levels of their learners, they are unable to use the formative assessment process to adapt their instruction to meet learning needs.

This pilot study aims to measure the efficacy of African home language formative assessment and evaluate the use of reading benchmarks in the improvement thereof. We do this by piloting a benchmarks-oriented intervention in schools across four provinces where the dominant language of learning and teaching is a part of the Nguni or Sesotho-Setswana language families. Using various metrics, we measure teacher knowledge of learning levels and estimate the intent-to-treat effect, local average treatment effect and treatment effect on the treated of the intervention on teacher knowledge of learning levels. Our findings inform our understanding of the basis on which benchmarks can be used productively in classrooms and hence contribute to the discussion of how benchmarks can be disseminated. The remainder of this paper is organized as follows: Section 2 provides background and context motivating the need to focus on assessment for learning in South Africa. Section 3 reviews the evidence on teacher judgement accuracy as an indicator of assessment efficacy and articulates the research questions explored in this study. Section 4 presents the early-grade reading benchmarks in African languages and provides details on the design of the intervention. Section 5 outlines the methodology and identification strategy. A descriptive analysis of the data is done in Section 6. Section 7 presents the results which are then discussed in Section 8. Finally, Section 9 concludes.

2. Background and Context

2.1. South Africa's 'Reading Crisis'

There is wide consensus that South Africa is facing a reading crisis and that learners are not mastering key foundational literacy skills during the foundation phase (Van der Berg et al.,

2011; Spaull, 2013; Taylor et al., 2018; Fleish & Dixon, 2019; Mohohlwane et al., 2022; Wills et al., 2022; Spaull, 2023). The 2021 Progress in International Reading Literature Study (PIRLS) results show that 81% of South African learners cannot read for meaning by the end of grade 4 (Department of Basic Education, 2023). Such studies have been important in quantifying the learning crisis and demonstrating that these learning gaps are formed in the foundation phase and continue to widen in later grades (Van der Berg, 2015). There is therefore a need to direct focus to addressing these gaps while learners are in the foundation phase to see sustained improvements in learner literacy outcomes (Spaull, 2023).

In response to South Africa's 'reading crisis', there are a range of ongoing initiatives and strategies to support early-grade reading. These include the provision of reading materials, campaigns to promote a culture of reading at school and home, improvements in initial teacher training and ongoing teacher professional development. These efforts rely on the buy-in of teachers and other school personnel to implement interventions. Teachers are a source of knowledge to learners and, in addition to parents and learners themselves, they are responsible for learning through implementing the curriculum. Their responsibility of educating learners and creating an environment for their cognitive development makes teachers prime candidates for agents through which stakeholders can intervene to improve learning outcomes.

In South Africa and elsewhere on the continent, there has been a suite of interventions referred to as 'Triple Cocktail' education interventions which are increasingly finding interest due to their potential to positively shift learner outcomes (Cilliers et al., 2018; Fleisch, 2018; Piper et al., 2018; Spaull, 2023). These are programs that include the provisions of scripted lesson plans, learning support materials and teacher training or support on effective pedagogic practice (Motilal & Fleish, 2020; Piper et al., 2018). Much of the focus of the teacher support element of these programs has been on teaching practice with little focus on other aspects of teaching like assessment practice.

More recently, the Department of Basic Education (DBE) has been leading efforts to establish language-specific benchmarks for foundational reading skills in all of South Africa's official languages. While reading benchmarks have long existed in English, these benchmarks cannot simply be transferred to all South African languages due to phonological, morphological, and orthographic differences between languages. To date, reading benchmarks have been established for the Nguni, Sesotho-Setswana language groups, Afrikaans and English First Additional Language (Mohohlwane, Wills & Ardington, 2022). Efforts to establish benchmarks in other languages are ongoing.

These benchmarks serve three primary purposes (Table 1). Firstly, they provide a national and provincial definition of proficient reading, allowing for target setting and monitoring of standards. Secondly, benchmarks are similarly useful at the school level to set targets, but more specifically to be able to standardise assessment practices and identify remediation needs. Lastly, benchmarks provide targets for teachers and learners at the classroom level, helping teachers establish criteria for successful progress and identify early on which learners are at risk of not learning to read for meaning by the end of the Foundation Phase. Additionally, benchmarks facilitate the interpretation of learners' assessment results, enhancing teachers' understanding of individual reading proficiency levels. This knowledge is essential for implementing the national curriculum's components that group learners according to their learning levels, such as group-guided reading. Moreover, effective remediation and consolidation for learners not meeting curriculum demands require teachers to have a solid grasp of both curriculum expectations and their learners' learning levels.

Table 1: Usage of Early Grade Reading Benchmarks

NATIONAL AND PROVINCIAL ADMINISTRATION	SCHOOL	CLASSROOM
Establishes definition of reading proficiency	Standards and targets that school leaders can aim towards	Standard against which to measure learner skills
Clearly communicates standards and targets	Standardises assessment practices across and within schools	Identify early on learners at risk of not being able to read
Monitor progress	Identify the extent of remedial support required	Adapt instructional focus to meet learners' needs

Source: Adapted from Ardington et al. (2020)

The classroom purposes of benchmarks to identify learners who are at risk of not being able to read and adapt instruction to meet learner needs relies on the quality of teacher formative assessment practices. Formative assessment is the mechanism through which teachers become aware of learning levels in relation to curriculum expectations and hence identify learners who are at risk. This information is then used to adapt their pedagogic practice to bring learners closer to where they should be (Kanjee & Bhanda, 2022; William & Thompson, 2007). The quality of assessment practices is often assumed and hence efforts to support

teachers are mainly directed at strengthening their teaching practice. For benchmarks to be effective in practice, the validity of this assumption needs to be explored further.

2.1. Formative Assessment Practices in South Africa

Existing research shows that there are inefficiencies in the formative assessment process that relate to teachers understanding of the purpose of formative assessment, how to interpret assessment results into learning needs and how to adapt their teaching practice. In a study to explore the assessment literacy of foundation phase teachers in three schools in Gauteng, Kanjee and Mthembu (2015) found that only half of the teachers could use formative assessment to identify learning gaps but did not know how to address them. Similarly, Mkhwanazi et al. (2014) considered how Siswati home language teachers in four schools use formative assessment in their home language teaching practice and found that teachers were unable to use formative assessment to support learning. Specifically, teachers provide learners with poor-quality feedback and only acknowledge the summative function of assessment which is to evaluate whether the learner has mastered a particular skill. In a larger sample study of the assessment practices in 54 South African schools, Kanjee (2020) found that teachers dominated the engagements where they were meant to gather evidence of learning. When learners are excluded from engagement during formative assessment, teachers do not obtain the information they need to adapt their instruction in order to enhance learning (Kanjee, 2020). The consequence of poor formative assessment practices is therefore that learning needs may not be addressed. In the first instance, this is because teachers simply do not have information on what those learning needs are. In the second instance, learning needs may not be met because teachers do not know how to adapt their instruction to address those needs.

There are several explanations for the lack of understanding of formative assessment and poor assessment practices. Firstly, there is a limited policy focus on assessment for learning, a lack of guidelines for implementing assessment for learning and a lack of capacity building for teachers in this regard (Kanjee & Croft, 2012; Kanjee & Sayed, 2013). The National Protocol for Assessment Grades R – 3 does not provide clear guidelines on how to use assessment to improve teaching practice despite mandating teachers to use it *"to provide feedback to the learners and teachers, close the gaps in learners' knowledge and skills and improve teaching"* (Department of Basic Education, 2011b). As a consequence, teachers prioritise recording and reporting over the effective use of assessment in teaching and learning (Kanjee & Croft, 2012). Secondly, teachers do not have access to sufficient reading resources

to assess learners and teach reading comprehension (Mkhwanazi et al., 2014). Teachers need access to text as an assessment resource but also to promote the culture of reading within their classrooms (Spaull & Pretorius, 2022). Lastly, the dominant form of engagement is teacher-led and collectivised which leaves limited opportunities for individualised assessment and feedback. This is mainly due to large class sizes which make assessment and the application of effective pedagogies like group-guided reading very difficult to implement (Cilliers et al., 2018; Kanjee, 2020; Spaull & Pretorius, 2022).

In response to these challenges with teacher assessment practice, Kanjee and Bhana (2022) evaluate the impact of a teacher professional development programme designed to improve teaching practice through the effective use of formative assessment. They found that the program was successful in improving teacher knowledge of formative assessment. They find some evidence that treatment teachers applied formative assessment strategies in their classrooms, like randomly selecting learners to answer questions and communicating the success criteria and learning criteria with learners. By randomly selecting learners to answer questions, teachers have a better sense of whether individual learners are grasping the content than for example if teachers asked the entire class to chorus the answer together. But teachers also need a more systematic way to obtain and record this information so that they are aware of individuals' abilities across the whole class.

To enhance formative assessment practices in reading, the DBE has been rolling out the Early Grade Reading Assessment (EGRA) to primary schools. This assessment tool, developed in 2006, is adaptable to multiple languages, cost-effective, and designed for easy administration by teachers or fieldworkers without specialized knowledge. In South Africa, the EGRA was versioned into all 11 official languages starting in 2007 (Govender & Hugo, 2020). The tool has been progressively introduced to primary schools through training led by Subject Advisors (Department of Basic Education, 2019b, as cited in Govender & Hugo, 2020). However, its implementation is not yet widespread, and the EGRA primarily serves as a classroom resource for teachers, with limited data collation and comparison within provinces (Mohohlwane, Wills & Ardington, 2022).

The EGRA is orally administered and assesses individual foundational reading skills, including the alphabetic principle, phonemic awareness, word recognition, fluency, and comprehension (Mohohlwane, Wills & Ardington, 2022). Each sub-task requires that a learner read as many items as possible from a chart of letters, syllables, words or connected text within one minute. The assessment takes approximately 15 minutes per learner and at the end, an assessor will have a score or correct letters, syllables or words read for each of the sub-tasks respectively.

At the time of the EGRA's development, African Language of Learning and Teaching (LoLT) reading benchmarks were not yet available. There was therefore no reading norm established of how teachers should interpret learner scores in the EGRA sub-tasks (Mohohlwane et al., 2020). Without these standards, teachers could not make informed normative judgements of whether learners were on track to being able to read with meaning by the end of Grade 3 based on their EGRA scores. The newly established African language reading benchmarks improve upon this by being primarily evidence-based, incorporating expert knowledge, and allowing for language-specific interpretations of reading scores. Where a learner's EGRA score is below the benchmark, this indicates that the learner is not on track and teachers can intervene through their instruction or remediation to bring them closer to the reading benchmark.

In addition to the lack of language-specific benchmarks, administration of the EGRA has proved challenging for teachers, particularly in the context of large classes. Teachers need text to administer the EGRA as an unprepared one-on-one oral assessment. Without text resources, learners are able to memorize the assessment passage and not demonstrate their true reading abilities. The assessment also involves mental math for score adjustments, managing multiple documents, and interpreting lengthy instructions. In large classes, individually assessing each learner consumes valuable time, making it difficult to use the EGRA frequently as a formative assessment tool. Revisions to the EGRA that reduce the cognitive burden on teachers and assessment time have the potential to enhance its usage for monitoring reading proficiency progress during the Foundation Phase.

While providing teachers with a user-friendly standardized assessment tool is important in strengthening assessment practices, it is not enough to ensure that assessment is used for improving learning. Teachers need to identify learning needs of learners in relation to curriculum demands and be willing and able to adapt instruction to meet those needs (Brunner et al., 2013). There are three key steps in the formative assessment process: 1) Establish where learners are in their learning, 2) Establish how far they are from where they need to be and 3) Establish what needs to be done to get them to where they need to be (William & Thompson, 2007). If teachers continue to teach at the pace of the curriculum despite formative assessment revealing that learners have not grasped foundational concepts, then simply conducting assessments will be ineffective at improving learning outcomes. Tailoring instruction is a key component of meeting identified learning needs.

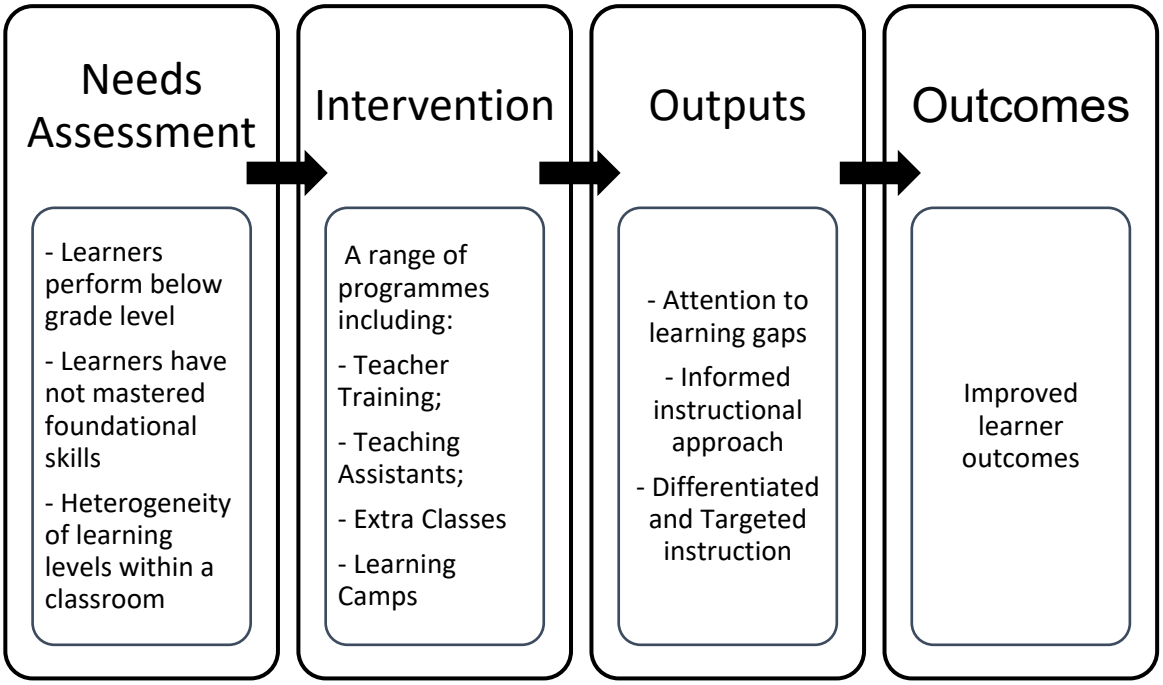
2.2. Differentiated Instruction

Effective assessment practice is crucial in tailoring instruction as it provides teachers with information on how to pitch instruction and provides learners feedback on how they can improve (Moon, 2005). Not only is tailored instruction needed to bring learners closer to where they should be, but it is also a requirement of the curriculum for language through Group Guided Reading. This is a reading teaching strategy where learners are divided into homogenous ability groups and provided with opportunities for individualised reading of an appropriately levelled text (Department of Basic Education, 2011a). In recent years, there has been growing interest in these kinds of differentiated instruction strategies to address low learner performance in many developing countries. Many learners struggle to keep up with curriculum demands, falling further behind as teachers progress through the curriculum without considering their understanding.

The Teaching at the Right Level (TaRL) approach, pioneered by the Indian NGO Pratham, targets instruction to learners' knowledge level rather than their grade level specified by the curriculum. Typically, these programs group learners by learning level for a portion of the school day or provide extra support after school for weaker learners. Such approaches have been found effective in improving learner performance, especially for at-risk learners (Banerjee et al., 2007; Bassi et al., 2020). In order to form ability groups or to target instruction, teachers need knowledge on student levels of attainment and their instructional needs in addition to subject-content knowledge (van Geel et al., 2019).

Figure 1 provides a typical theory of change for targeted instruction programs, with each link in the chain underpinned by assumptions that enable the causal pathway from inputs to outputs to outcomes. For example, the output “differentiated and targeted instruction” assumes that teachers i) know the level of their learners, ii) understand the needs associated with each level, iii) understand how to address those needs and iv) are willing and able to do so. If any of these assumptions are not met, the causal chain is broken.

Figure 1: Typical Theory of Change of Targeted Instruction Programs



These assumptions are typically taken as given when in fact some are quite strong. For example, knowledge of learning levels. The evidence around formative assessment practices and the quality thereof in South Africa makes it difficult to argue that teachers actually know the learning levels of their learners. This assumption in particular requires further interrogation and is the focus of this study.

3. Present Study

The first step in the formative assessment process is to establish the learning levels of learners. Teachers need accurate information on students' learning levels in order to form knowledge on student abilities which they can use to make decisions related to their teaching practice.

3.1. Previous research on Teacher Judgement Accuracy

One way the quality of this knowledge can be evaluated is through measuring the consistency of teacher judgements with objective assessments of learners' academic abilities (Kaufmann, 2020; Kolovou et al., 2021). This is widely referred to as teacher judgement accuracy and assuming that objective learner assessments are reliable and valid, it is a proxy for teacher knowledge of learning levels. Teacher judgement accuracy is widely operationalised as the correlation between teacher judgements and learner performance (Kolovu et al., 2021). To this

end, teacher judgement accuracy has been found to be low (Hadjidemetriou & Williams, 2001; Maunganidze et al., 2008; Kilday et al., 2012; Djaker et al., 2022).

In a meta-analysis of 75 studies from developed country contexts (for an exception from a developing context, see (Maunganidze et al., 2008)), Sudkamp et al. (2012) found an average correlation of 0.63 between teacher estimates and actual learner performance. This corresponds to a moderate positive correlation but in the context of well-functioning assessment systems, this is relatively low (van der Berg & Shepherd, 2015). In a similar manner, Kaufmann (2020) re-analyse the meta-review data of 16 studies in Hoge and Coladarci (1989) and found an average correlation of 0.8 whereas the original study found an average correlation of 0.65. They argue that the original study only corrected for sampling error but not measurement error (reliability of assessment) and loss of information from dichotomising a continuous variable therefore it would have underestimated teacher judgement accuracy. Kolovou et al. (2021) also considered methodological issues of teacher judgement accuracy and argued that there are limitations related to sampling and measurement error that may attenuate correlation measures. Instead, they make use of a multivariate multi-level latent modelling approach and find evidence of low to medium average teacher judgement accuracy in mathematics and language. Interestingly, they also find that teacher judgement accuracy is not content-specific within a specific subject. For example, if a teacher is able to accurately judge a learner's letter sound knowledge, they would also be able to accurately judge a learner's oral reading fluency proficiency.

The contentions around teacher judgement accuracy and the measurement thereof present an opportunity for further research. The body of literature in this regard is growing but has at present primarily been from studies in more developed contexts where research has progressed beyond measurement to exploring factors affecting teacher judgement accuracy. Some studies investigate whether teacher judgement accuracy can be explained by bias towards certain student demographic characteristics. For example, Ready and Chu (2015) investigate the presence of Pygmalion effects in teacher judgement accuracy by exploring the links between teacher estimates and student performance in the United States. They find teachers overestimate the performance of higher socio-economic status (SES) learners than they do their lower SES counterparts.

The literature on teacher judgement accuracy in developing countries is scarce. Djaker et al. (2022) evaluated the judgements of primary and middle school teachers in India and Bangladesh against learner test performance and found that teachers misestimate the performance of their learners by a large margin and in particular that they tend to overestimate

the performance of their learners. They further find that poor teacher judgement accuracy could not be explained by teacher characteristics such as educational background, years of experience or teacher training. In a small sample study in Zimbabwe, Maunganidze et al. (2008) found a zero correlation between teacher estimates and actual learner performance in mathematics and a 0.36 correlation in reading.

In South Africa, there have only been a few studies that look at teacher judgements of learner performance and none of which were of teachers in the foundation phase. Gamaroff (1999) does not specifically look at the accuracy but rather the inter-teacher reliability when teachers mark the same set of essays and found large variation in how teachers rated the essays. Van der Berg and Shepherd (2015) on the other hand compare school-based continuous assessment results with externally moderated assessment results of the same learners and find a weak correlation between the two and specifically that teachers overestimated the performance of learners. They find further that teachers did not use the feedback from external examinations to re-evaluate their assessment practices which caused the gap in assessment results to widen over time. Given the findings of prior research on poor formative assessment practices, it is unsurprising that teachers are unable to reliably estimate the performance of their learners. This suggests that teachers may not be aware of the learning levels of their learners.

3.2. Research aims and questions

The overarching purpose of this study is to generate insights into how the newly established benchmarks could be productively used in South African classrooms. We focus on the intended uses of the benchmarks which are to identify learners who are at risk of not being able to read and adapt instruction to meet learner needs. We do this by examining current teacher knowledge of learning levels through their existing formative assessment practices in literacy and evaluating the effectiveness of a light-touch intervention in improving that knowledge. Using benchmarks to improve teacher knowledge of learning levels creates a basis for adapting and tailoring instruction to meet learning needs and improve learning outcomes.

We therefore define our research questions as follows:

1. What is the level of teacher judgement accuracy?
2. What is the effect of providing training on benchmarks and assessment resources on teacher judgement accuracy?

3.3. Contribution to literature

Several gaps emerge in the literature which the present study aims to contribute to. The first is that amongst the literature on formative assessment practice in South Africa, the majority are based on sample studies of three to four schools and where there were fairly large sample size studies, these did not focus specifically on the foundation phase and evaluated the use of specific assessment strategies (see Kanjee (2020) and Kanjee & Bhana (2022)). While these studies are helpful to gain insight into teachers' use of these strategies, the lack of focus on the foundation phase, where learning gaps form, means we have little insight into how benchmarks could be used productively in this context. Secondly, there has been no study to date that evaluates the in-classroom use of the newly developed reading benchmarks. These benchmarks are an important policy development towards tangibly quantifying what is meant by "reading for meaning" but the extent to and manner in which they can be operationalised has not been empirically tested. The present study hopes to contribute to closing the gap between policy and implementation. Finally, there is very little evidence on whether South African teachers can accurately judge the performance of their learners. This skill is often assumed, arguably incorrectly, and hence overlooked despite it being vital in teachers implementing assessment for learning and targeted instruction aspects of the national curriculum like group-guided reading. The present study hopes to fill this gap by firstly measuring teacher judgement accuracy and secondly, evaluating an intervention designed to improve this accuracy. In the next section, we describe the intervention before setting out our methodological approach to answering our research questions.

4. Intervention

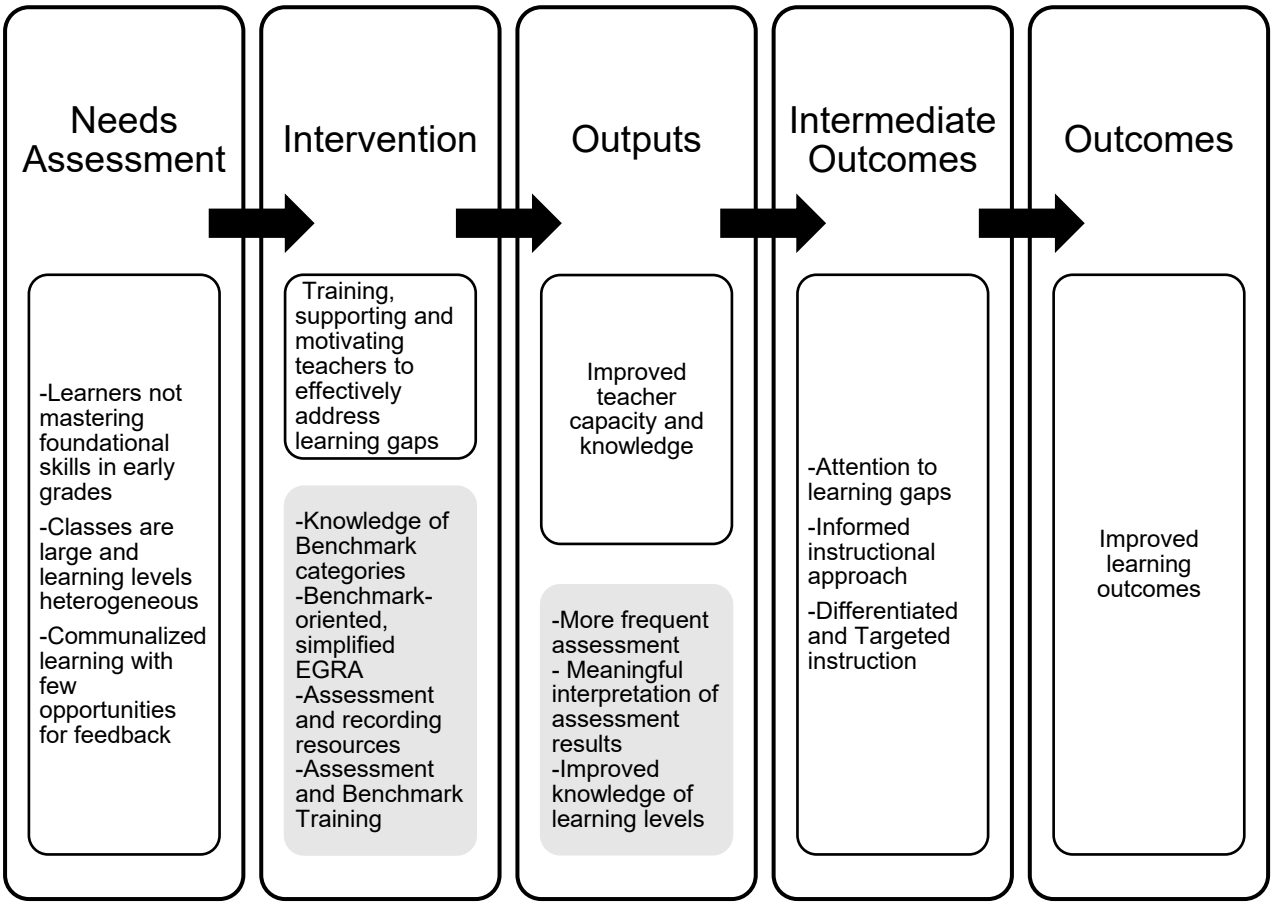
The intervention was designed to improve the efficacy of formative assessment in early-grade reading. This is achieved by introducing teachers to the newly established early grade reading benchmarks for the foundation phase LoLT in their school, simplifying the administration of the EGRA and linking the EGRA results to the benchmarks. Through clearly articulating curriculum expectations with respect to reading fluency; reducing the burden of conducting the EGRA; facilitating the meaningful interpretation of the EGRA results and providing an accessible visual summary of learner progress, the intervention sought to improve teacher judgement accuracy.

We revisit a more detailed version of the theory of change shown in Figure 1 to locate this intervention within the broader aim of improving foundational literacy outcomes.

4.1.1. Theory of Change

Our theory of change relating to the intervention is outlined in Figure 2.

Figure 2: Theory of Change



Learners are performing below grade level in reading. However, due to challenges related to assessment and result interpretation, teachers may lack a clear understanding of the scope and depth of these learning gaps. This is primarily because they are unaware of their learners' specific reading abilities in absolute terms and in relation to curriculum expectations. This issue is exacerbated by the context in which South African teachers teach; class sizes are large and learners' reading proficiency is heterogeneous. South African classrooms are characterised by communalized learning with few opportunities for individualised feedback or targeted instruction (Hoadley, 2018).

In a broader sense, when teachers differentiate instruction as required by the curriculum, they provide specialized instruction to learners based on their ability levels. They cannot begin to differentiate if they do not know what those levels are. Knowing the learning levels of learners is just the first step; it is not enough on its own. Teachers also need to understand the unique learning needs of each level and have strategies to address those needs. Nevertheless, knowing the learning levels of their students is an essential prerequisite for teachers to successfully implement differentiated instruction and ultimately improve learning outcomes. Prior to the benchmarks, teachers also did not have an informed standard of what reading milestones learners in their respective grades needed to meet.

The goal of any education intervention is to ultimately improve learning outcomes; in this case reading proficiency. Given contextual factors like learners performing below grade level, large class sizes, heterogeneous learning levels and communalized learning, differentiated instruction has demonstrated the potential to improve learning outcomes. For differentiated instruction to be effective in improving reading abilities, teachers need to know the learning levels of learners. By introducing benchmarks, simplifying the EGRA, and providing assessment materials and training, the intervention will improve the efficacy of formative assessment. With this intervention, teachers will assess more frequently and interpret the results of that assessment more effectively. This will create more opportunities for teachers to engage with the reading levels of their learners and provide them with a metric against which to appropriately define reading proficiency. As a result, teachers will improve their judgement accuracy. This, in turn, will enable teachers to use their accurate judgements of learning levels in addressing learning needs through differentiated instruction. This link is highlighted as the light grey blocks in the theory of change diagram in Figure 2 and is also the focus of this intervention. This is not all that is required for teachers to be able to differentiate their instruction – they also need training and support on how to differentiate their instruction to address the learning needs associated with each level. However, this step can only happen if teachers are aware of what those learning levels are.

4.2. Intervention Design

4.2.1. Early Grade Reading Benchmarks

The newly established benchmarks for the African languages included in this study are shown in Table 2.

Table 2: Early Grade Reading Benchmarks for Nguni and Sesotho-Setswana Languages

Language Group	Grade 1	Grade 2	Grade 3
	By the end of the end the year, learners should be able to:		
Nguni	sound out **40 letters correctly** in **one minute**	**correctly read 20 words** in a **passage**.	**correctly read 35 words** in a **passage**.
Sesotho-Setswana		**correctly read 40 words** in a **passage**.	**correctly read 60 words** in a **passage**.

The benchmarks were introduced to teachers in colour-coded categories as shown in Table 3 and Table 4. This layout breaks down the benchmarks into five categories of proficiency: non-reader, struggling reader, emerging reader, proficient reader, and in the instance of Grade 2, fluent reader. The rationale for this breakdown is that since the benchmarks are an 'end-of-year' metric, they are not useful to measure progress within the grade. The categories are helpful for teachers to plot whether their learners are on track to attaining the target at the end of the year as the terms progress and how many words or letters they need to gain to meet the benchmark. The naming convention is also a good way to describe how learners are faring compared to the expectation. The proficient reader category is the minimum benchmark learners need to reach by the end of the year, whereas in the Grade 2 instances where there is a category for 'fluent reader', it means that the learner is reading at the proficiency level of the next grade.

Table 3: Letter Sound Knowledge Benchmark Categories

Non-Reader	**0**
Struggling Reader	**1-25**
Emerging Reader	**26-39**
Proficient Reader	**40 and above**

Table 4: Oral Reading Fluency Benchmark Categories

NGUNI LANGUAGES

Category	Grade 2	Grade 3
Non-Reader	0	0
Struggling Reader	1-9	1-19
Emerging Reader	10-19	20-34
Proficient Reader	20-34	35 and above
Fluent Reader	35 and above	

SESOTHO-SETSWANA LANGUAGES

Category	Grade 2	Grade 3
Non-Reader	0	0
Struggling Reader	1-19	1-39
Emerging Reader	20-39	40-59
Proficient Reader	40-59	60 and above
Fluent Reader	60 and above	

4.2.2. Enhanced EGRA Assessment Tool

Teachers attended a full-day training which was conducted by the Department of Basic Education. This training was done in geographic clusters of schools. Teachers were introduced to the study, the research aims and trained on how to classify each learner's reading proficiency according to the benchmark categories using an EGRA-like assessment. The adapted EGRA tool for this intervention was restricted to the two skills that match the benchmarks: Letter Sound Knowledge and Oral Reading Fluency. By reducing the sub-tasks to be administered from four to two, the time per learner including giving instructions was reduced from up to 15 minutes (Hugo & Govender, 2020) to five minutes. This means that a class of 45 learners will take just under four hours.

To implement the intervention, teachers received a teachers' handbook explaining the different colour codes explained in Table 3 and Table 4 as well as two identical letter sound and ORF assessment charts. The teacher version of the assessment chart was colour-coded denoting the reading categories as seen in Figure 3 and Figure 4 while the version for use by learners was not colour-coded to minimize distractions to the learners. The teacher assessment chart was wipeable allowing teachers to use the same chart for every learner. In our adapted EGRA tool, learners are instructed to sound out as many letters or read as many words as they can on the respective charts in 60 seconds for each task. The teacher starts the timer on their cellphone or stopwatch as the learner begins sounding out or reading and crosses off any letter or word the learner gets incorrect on the teacher chart. At the end of the 60 seconds, the teacher marks the final attempted letter or word and then for each incorrect letter or word, the teacher moves one letter or word backwards on the chart. The colour associated with this final position on the chart determines the benchmark category for each learner.

Traditional EGRAs require teachers to calculate an exact fluency measured in correct letters per minute (CLPM) or correct words per minute (CWPM). To do this, the teacher needs to count the number of attempted items and incorrect items or subtract the latter from the former. In instances where learners complete the task before 60 seconds, the teacher is required to record how many seconds are remaining and then adjust the number of correct items upwards to account for the shorter time. In our modified EGRA, teachers only need to keep track of incorrect items and record a colour, reducing the cognitive burden of assessment.

Our modified EGRA not only simplifies the administration of the assessment but also provides a ready interpretation of the result. The benchmarks were developed based on empirical evidence to be language-specific and in line with the theory of fluency development. The categories therefore correspond to distinct and meaningful stages in a learner's reading development.

Figure 3: Letter Sound Knowledge Chart

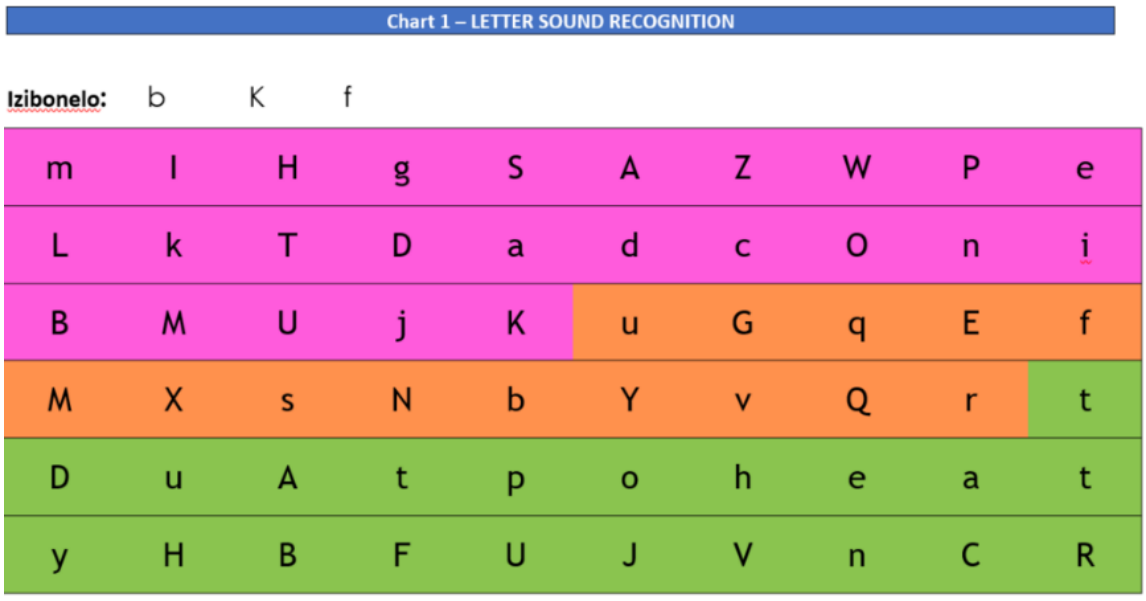
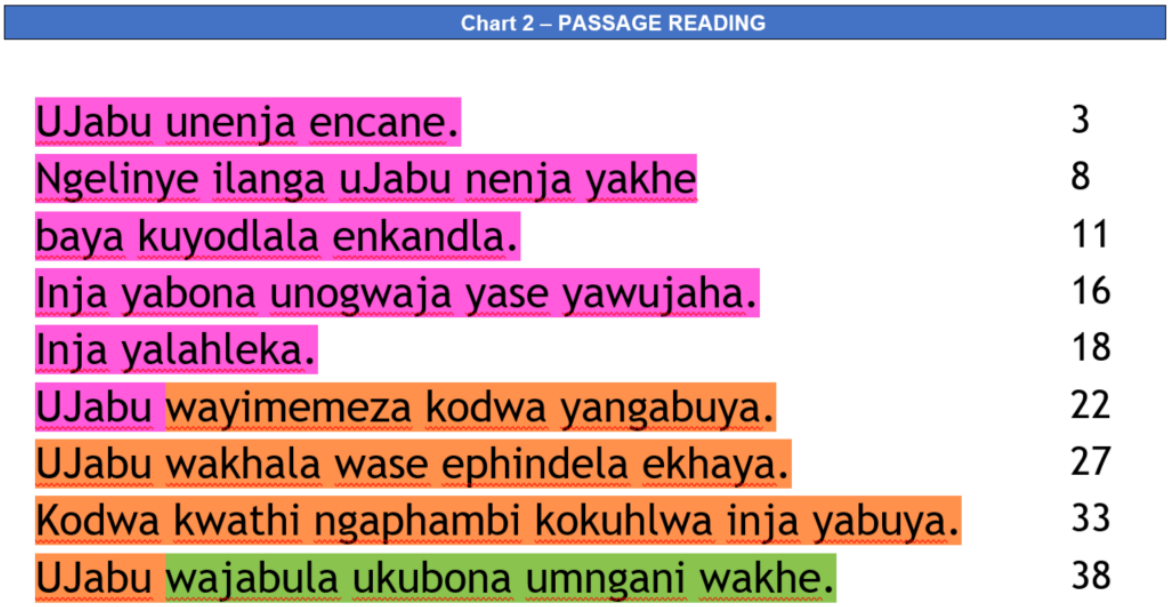
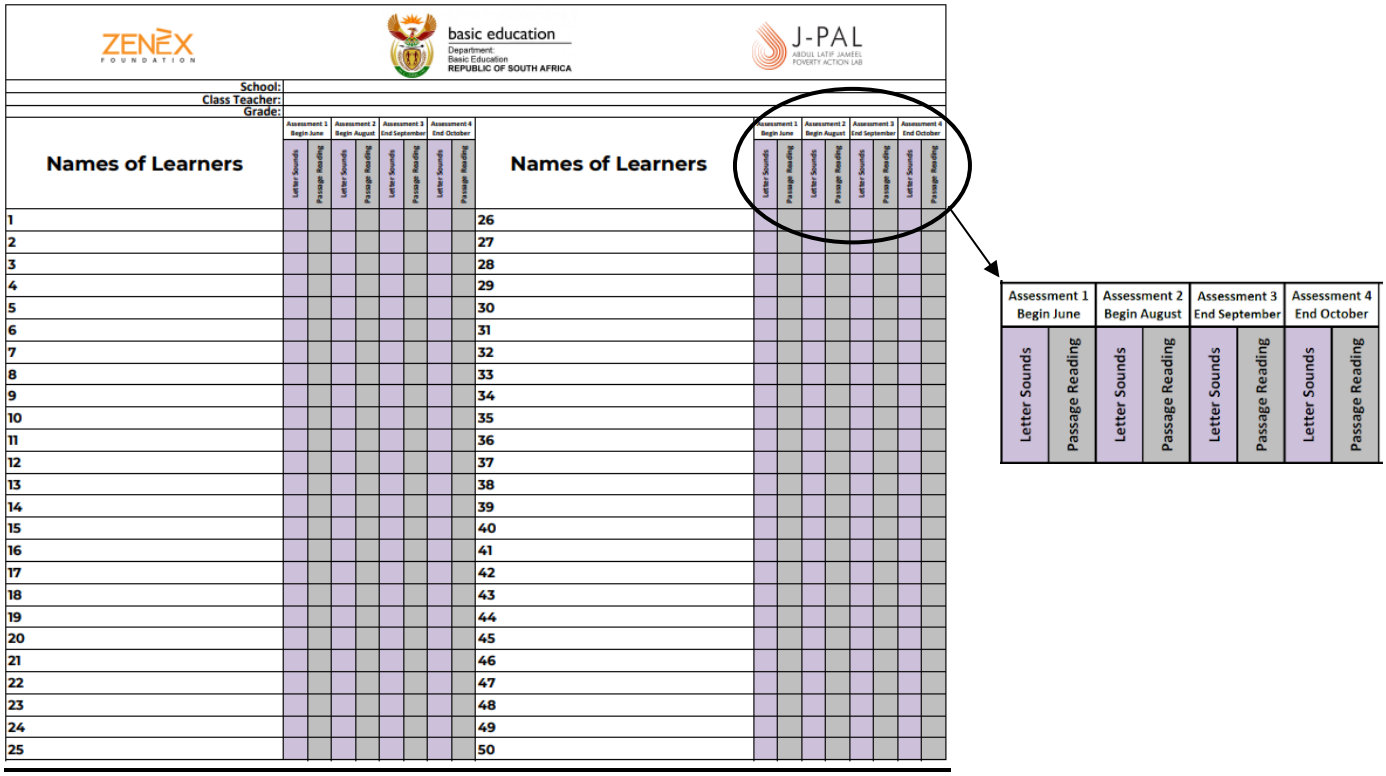


Figure 4: isiZulu ORF Passage Teacher Chart



The final enhancement to the EGRA was the class progress chart shown in Figure 5 together with stickers matching the reading level colour codes. After assessing each learner, the teacher was required to place the relevant colour sticker next to the learner's name and task on the progress chart that is displayed in the classroom instead of recording a number in a booklet. This allows easy access for the teacher to see at a glance where each learner is placed. At the end of each assessment period, the teacher would be able to use the information on the charts to intervene and adapt instruction to meet their learners' needs. Over time, the chart would reveal the learners' progress toward reaching the benchmarks by the end of the academic year.

Figure 5: Class Progress Chart



ZENEX FOUNDATION

basic education
Department: Basic Education
REPUBLIC OF SOUTH AFRICA

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

School:
Class Teacher:
Grade:

Names of Learners	Assessment 1 Begin June		Assessment 2 Begin August		Assessment 3 End September		Assessment 4 End October		Names of Learners	Assessment 1 Begin June		Assessment 2 Begin August		Assessment 3 End September		Assessment 4 End October	
	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading		Letter Sounds	Passage Reading	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading
1									26								
2									27								
3									28								
4									29								
5									30								
6									31								
7									32								
8									33								
9									34								
10									35								
11									36								
12									37								
13									38								
14									39								
15									40								
16									41								
17									42								
18									43								
19									44								
20									45								
21									46								
22									47								
23									48								
24									49								
25									50								

Assessment 1 Begin June		Assessment 2 Begin August		Assessment 3 End September		Assessment 4 End October	
Letter Sounds	Passage Reading	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading	Letter Sounds	Passage Reading

Teachers were asked to assess learners four times between June and October on Letter Sound Knowledge and Oral Reading Fluency using the intervention materials they received. According to the Curriculum and Assessment Policy Statement (CAPS), a minimum of seven hours is allocated for home language learning per week of which 30 minutes per day (2.5 hours per week) are set aside for group guided reading. Teachers were advised to use two weeks of group guided reading time to conduct each round of assessment. The teacher used the same Letter Sound Knowledge and Oral Reading Fluency charts at all four assessment points. It was explained to teachers that they were not to use the charts for any purpose other than during the four assessment periods to avoid learners committing the charts to memory and therefore not displaying their true reading ability but rather their ability to memorize.

4.2.3. Differentiated Instruction

We further provide high-level guidance on the appropriate use of the benchmark categories in the classroom over and above the assessment component. These include identifying learners who are at risk of not learning to read by the end of the foundation phase; which are learners whose reading proficiency falls below a certain threshold from the benchmark. We suggest that teachers group learners for activities such as group guided reading according to the

benchmark categories and targeting instruction at the level of the learners individually and across the various ability groups.

4.2.4. Resources for Teaching

Teachers were encouraged to lean on their expert knowledge on how to address learner needs for each ability level. In addition to this, teachers were directed to the National Framework for the Teaching of Reading in African Languages in the Foundation Phase. This document outlines the different skills that build up to fluent reading and comprehension; how to identify gaps in these skills; and how to remediate them. Another resource that was made known to the intervention teachers was an online teacher development platform[1] that offers teachers courses that help strengthen pedagogical practice including the different subsets in reading instruction.

5. Methodology

In this section, we describe the methodological approach to answering the research questions.

5.1. Sample Selection

At the time of the study, the only reading benchmarks that were developed and publicly available were for the Nguni and Sotho-Setswana language families. We therefore chose provinces where the majority of schools had one of these languages as the language of learning and teaching (LoLT). Due to the modest scale of the pilot study, we also wanted to minimize the geographical scope of the study sites and opted for Eastern Cape for isiXhosa, Mpumalanga for isiZulu, North West for Setswana and Limpopo for Sepedi. Our sample is restricted to Quintile 1-3[2] schools in urban areas which serve communities of lower socio-economic status and schools with at least three classes per grade. Working with the Department of Basic Education, we selected schools that had received training on EGRA. The intention behind using the EGRA database was to build on and refine the existing learner assessment practices amongst these schools rather than to introduce early-grade reading assessments. From this database, we stratified by province to randomly draw a sample of 15 schools per province with the intention that 10 schools would be evaluation schools and five would be reserve schools in case a school declined to participate. Within each school, we randomly selected three teachers in Grade 2 and three teachers in Grade 3 to participate in

[1] www.tpd-dbe.org
[2] In South Africa, public schools are organized into 5 quintiles ranging from Quintile 1 (located in the poorest communities) to Quintile 5 (located in the least poor communities).

the study. For each teacher, we also randomly selected 10 learners to assess. Our desired sample therefore consisted of 40 schools, 240 teachers and 2400 learners. Ethical clearance[3] for this study was obtained from the Commerce Ethics in Research Committee at the University of Cape Town

Upon visiting the schools during data collection, it was discovered that some schools did not have the required number of classes contrary to the information captured in the database. Specifically, none of the schools in North West and Limpopo had enough classes. The sampling frame was therefore extended to include schools that were not part of the EGRA project rollout. The non-EGRA schools were randomly sampled within each LoLT from the Education Management Information Systems (EMIS) database.

5.2. Instruments

We administered three different survey instruments. The first was a teacher questionnaire which contained questions on demographic characteristics, teaching experience, home language teaching practice and assessment practice. On assessment practice, teachers were asked about the frequency and implementation of EGRA, to estimate the number of words and letters each sampled learner could read within one minute and to identify the strongest and weakest readers among the 10 sampled learners. To measure cognitive burden, we also asked teachers questions on their attitude towards the EGRA and their beliefs on what they believe would make it easier to administer. The questions on easing administration were asked after the questions on attitude so as not to prime teachers that the EGRA may be difficult to administer. The second instrument was the school principal or HoD questionnaire which consisted of questions on the class sizes and number of teachers in the foundation phase as well as school resources.

The third instrument was the learner assessment in which we included sub-tasks of the EGRA. These were developed for the African languages through the work of numerous academics and education practitioners and have been used in previous education impact evaluations. A key advantage of using the EGRA is that it is widely used and understood, facilitating comparisons across a range of programmes. For consistency with the intervention and comparability, the assessment included the Letter Sound Recognition and Oral Reading Fluency sub-tasks. We also included comprehension questions related to the Oral Reading Fluency sub-task for informational purposes although this task was not included in the

[3] Commerce Ethics in Research Committee Application Reference: REC 2022/04/011

intervention materials. Learners were given one minute to sound as many letters or read as many words as possible and they were given an additional two minutes in the oral reading fluency tasks so that they were able to answer the comprehension questions.

5.3. Data Collection and Randomisation

We conducted school visits at two points: once prior to the introduction of the intervention to collect baseline data in May 2022 and again six months after the intervention was introduced in October and November 2022. At baseline, 228 teachers were successfully interviewed. Amongst these teachers, we then stratified by school and grade to randomly assign two teachers to the treatment group and one teacher to the control group. Schools were then informed of the intervention training and given the list of teachers who were invited to attend. Not all of these teachers attended the training. In some cases, schools swapped them out for teachers who had not been invited, in most instances at the discretion of the principal. Table 5 combines attendance register data from training with the invitation lists to show the number and proportion of those invited who attended.

Table 5: Training Invitation and Attendance

	Invited to training			**Not invited**		
Province	Attended	Did not attend	% Attending	Attended	Did not attend	% Attending
Eastern Cape	32	4	89%	4	16	20%
Limpopo	32	4	89%	5	14	26%
Mpumalanga	32	7	82%	1	18	5%
North-West	33	7	83%	3	16	16%
Total	**129**	**22**	85%	**13**	**64**	17%

Note: Table reports the number of teachers who attended training of those who were randomly selected to attended training and those who did not receive an invitation to training. N = 228

Eighty-five percent (85%) of teachers invited to the training attended with the lowest attendance rate in Mpumalanga province. In many instances where invited teachers did not attend, they were replaced by non-invited teachers from the same school. Overall, 17% of non-invited teachers in our sample attended the training. One of the 40 schools did not send any teachers to the training. For the descriptive analysis, we compare teachers who attended training (trained) to those who did not (untrained). We return later to consider fidelity to the intervention when we estimate the impact of the intervention regardless of whether teachers were invited or not to receive the training. In Sections 5.5 and 5.6 we set out different definitions of treatment and control groups for estimating the impact of the intervention.

At endline, 202 of 228 teachers were successfully re-interviewed. The main reasons some teachers were not re-interviewed were because they were absent during the fieldwork period or they had moved to a different school. Table 6 reports the number of teachers interviewed in each round by training status. The overall attrition rate is 12% and 10% in the trained group and untrained group respectively which suggests that differential attrition is not a problem. The sample for analysis is restricted to the 202 teachers observed at both data collection points. We further removed the five[4] teachers from the school where no teachers attended the training as we aim to have perfect balance in the schools across trained and untrained. The analytical sample therefore includes 197 teachers.

Table 6: Teacher Sample

	Baseline		Endline			
	Interviewed		Interviewed		Attrited	
Province	Trained	Untrained	Trained	Untrained	Trained	Untrained
Eastern Cape	36	20	31	18	5	2
Mpumalanga	33	25	30	22	3	3
North-West	37	18	32	16	5	2
Limpopo	36	23	32	21	4	2
Total	142	86	125	77	17	9

We also randomly selected 10 learners per teacher for assessment however, several issues emerged during baseline fieldwork which caused deviations from the original sampling plan. These included strike action blocking access to schools, disruptions to school water supply causing a shortened school day etc. As a result, there are a few cases where fewer than 10 learners were sampled per teacher and the total baseline sample size was 1,998. At endline, we re-assessed the same learners but where these learners were unavailable, they were replaced. The analytical sample was restricted to the learners of the 197 teachers who were observed at baseline and endline. Table 7 reports the learner sample size by grade and province.

[4] One of the six teachers in this school was not successfully re-interviewed at endline and thus does not form part of the 202.

Table 7: Learner Sample

	Grade 2		Grade 3	
Province	Baseline	Endline	Baseline	Endline
Eastern Cape	252	262	224	189
Mpumalanga	228	227	236	236
North-West	240	227	240	245
Limpopo	249	238	279	234
Total	**969**	**954**	**979**	**904**

5.4. Measuring Teacher Judgement Accuracy

The main outcome we wish to measure is teachers' knowledge of learning levels. We proxy knowledge as teacher judgement accuracy because as teachers' knowledge of learning levels improves, they should be more likely to accurately judge learning levels. There have been a variety of methods used in measuring teacher judgement accuracy in the prior literature. These include estimating the amount of variation in teacher judgements that can be explained by variation in learner performance (Djaker et al., 2022; Maungandize et al., 2008), comparing the difficulty rating of assessment tasks with the proportion of learners that correctly answered the question (Hadjimetrou & Williams, 2001; Brunner et al., 2013) and calculating the difference in teacher estimations and learner performance (Djaker et al., 2022; Brunner et al., 2013). However, teacher judgement accuracy has predominately been operationalised as the correlation between teacher estimates of learner performance and actual learner performance. One approach in this regard is to calculate the Pearson correlation between the scores that learners actually obtained on an external assessment and the teacher rating on the same scale (Maunganidze et al., 2008; Martinez et al., 2009; Sudkamp et al., 2012; van der Berg & Shepherd, 2015). A related approach is to estimate the correlation between teacher rankings of learner performance and the ranks of learners based on their assessment performance using Spearman's Rank Correlation (Brunner et al., 2013; Ready & Chu., 2015; Kolovou et al., 2021; Djaker et al., 2022). This measures the degree to which there is agreement in the rank orders of teachers' judgements and student performance and hence has been referred to as a measure of diagnostic sensitivity (Brunner et al., 2013).

There are several methodological challenges to estimating teacher judgement accuracy. The first is that measures of student performance may contain random error (Kaiser et al., 2017). Measuring student reading proficiency based on a single assessment conducted in an unfamiliar environment can be influenced by factors such as learners' comfort, learners being

able to read specific familiar words but none of these being included in assessment text etc. This would introduce measurement error into measures of learner performance as the assessment results may not reflect the learners' true reading abilities. A teacher who interacts with learners daily may be more familiar with learners' true academic abilities. However, in the presence of measurement error, their estimates of learner performance may not be strongly correlated with measured learner performance. This would bias measures of teacher judgement accuracy downwards not because the teacher has poor judgement, but because learner abilities were poorly measured.

Sudkamp et al. (2012) propose a model of four key factors that influence teacher judgement accuracy. The first factor is judgement characteristics where they find that teachers are more accurate when informed of the test on which their judgements are based, and when they estimate for a particular skill or content domain as opposed to general academic achievement. The second factor is the characteristics of the assessment with better accuracy associated with curriculum-based measures as opposed to standardized achievement tests. Curriculum-based assessments are more closely related to learner in-class performance and more aligned with the teaching environment. The third factor is teacher characteristics such as academic qualifications or teaching experience. The fourth factor is student characteristics where there may be bias in how teachers judge learner performance based on demographic characteristics. The first two factors present the second methodological challenge with measures of teachers' judgement accuracy dependent on the means by which such judgements are elicited. For example, teachers should be better able to assess performance for a particular skill, such as oral reading fluency, as opposed to estimating learner performance more generally, for example, 'reading proficiency'. If assessments are unseen or not closely aligned with the curriculum, estimates of teacher judgement would be a poor proxy of true judgement due to noise and hence measures of judgement accuracy would be attenuated.

The final limitation is that the correlation only measures the accuracy of teacher judgements of relative learner performance and not necessarily the absolute level of learner performance. For example, a teacher may correctly estimate that one learner performs better than another learner but could be completely overestimating the performance of both or one of those learners. In this case, the correlation would overestimate teacher judgement accuracy. As noted above, there is potential for measurement error in learner performance and noise in teacher estimates, when these two measures are correlated, the correlation would be biased downward. The correlation would underestimate teacher judgement accuracy. Secondly,

comparing rank orders of the two measures does not provide information on the precision of teacher judgements or whether teachers tend to over or underestimate learner performance.

Teachers knowing how to accurately rank learners' academic abilities may also not be useful when considering teacher judgement accuracy in the context of formative assessment practices. Learner rankings provide no direction to the teacher on how to differentiate instruction or to remediate. An alternative measure of teacher knowledge of learner relative reading proficiency is whether teachers are able to identify the strongest and weakest readers from a list of learners we randomly select from sampling. This measure is still crude and imperfect because it is still subject to noise due to how broadly teachers could interpret the "strongest" or "weakest" reader. Secondly, two teachers who correctly identify the strongest or weakest reader cannot be compared further even though the one may have a deeper knowledge and understanding of the reading proficiency of their learners. Finally, identifying the strongest or weakest learners is still not meaningful to teachers when using formative assessment to inform differentiation. Using measures of absolute learner performance in curriculum-aligned content domains in addition to correlation measures could be more meaningful for measuring teacher judgment accuracy and using that accuracy to adapt instruction.

To address the measurement error in learner performance, we employed a team of highly qualified fieldworkers with experience in conducting learner assessments with young children. Learners were also informed that the assessment was low-stakes and confidential in order to make them comfortable with reading and build rapport. The field team was therefore well-suited to create an environment where test anxiety was minimized, and learners displayed their true reading skills.

To minimise noise in teacher estimations, we ensure that teachers make informed judgements by asking them to estimate learner performance on the same test materials presented to learners (letter sounds chart and passage of connected text). We also focus on two specific content domains which are the number of correct words read and the number of correct letters read and ask teachers to estimate these separately as opposed to providing an overall rating of reading proficiency. In focusing on these two content domains, we also make sure that we are able to align teacher judgements to curriculum-based measures, namely the newly established reading benchmarks. We do not ask teachers explicitly to rank the learners as we cannot account for other judgement-related factors that go into such an estimation. For example, whether a teacher would be factoring in general academic performance outside of home language literacy. Rather, we calculate learner rankings according to the teacher directly

from the teacher point estimates of number of correct words or letters. These point estimates are based on a well-defined and specific criterion in terms of number of words or letters that leaves little room for teachers to account for unmeasured factors. Due to budget constraints and the small scope of the pilot study, we did not collect any information on student demographic characteristics and therefore do not directly control for these. We do however randomly select learners using the lottery method in order to create a representative sample of the teacher's class. Teachers' estimations of the performance of this learner sample should be correlated to their judgements under normal in-class conditions. We discuss how we account for teacher characteristics in the model specification section below.

In response to the final limitation that correlation is an insufficient measure on its own of overall accuracy, we use multiple measures in addition to the correlation. Firstly, we define diagnostic sensitivity as the correlation between the implied teacher rankings and the rankings of learners based on the assessment scores. Following Brunner et al. (2013) we define judgement error as the absolute difference in the teacher estimates and learner performance as a measure of judgement precision. We derive the teacher estimate of a learner's benchmark category based off of their point estimate. The accuracy of teacher predictions of learner benchmark categories can be thought of as a classification problem which can be investigated by estimating the number of predictions that teachers correctly classify. We borrow from the machine learning literature to construct a multi-class confusion matrix (Grandini, Bagli & Visani, 2020) and measure teacher judgement accuracy using the accuracy rate metric defined as follows:

$$TP_{it} = \sum_{x=1}^{5} \sum_{j=1}^{10} 1[y_{ijt} = x;\ \hat{y}_{ijt} = x] \quad (1)$$

$$TN_{it} = \sum_{x=1}^{5} \sum_{j=1}^{10} 1[y_{ijt} \neq x;\ \hat{y}_{ijt} \neq x] \quad (2)$$

$$FP_{it} = \sum_{x=1}^{5} \sum_{j=1}^{10} 1[y_{ijt} \neq x;\ \hat{y}_{ijt} = x] \quad (3)$$

$$FN_{it} = \sum_{x=1}^{5} \sum_{j=1}^{10} 1[y_{ijt} = x;\ \hat{y}_{ijt} \neq x] \quad (4)$$

Where y_{ijt} is the measured benchmark category (x) of learner j in teacher i's class in time t. $\hat{y}_{ijt}$ is the teacher predicted benchmark category of each learner. TP_{it} is the number of predictions, summing over learners and categories, where a teacher predicts the reference category correctly. For a learner who is classified as a struggling reader, this value would be equal to one when the reference category is struggling reader, and the teacher estimates that the learner falls into the struggling reader category. TN_{it} is the number of predictions where a teacher correctly predicts that a learner does not fall into the reference benchmark category. For example, when the reference category is proficient reader, but the learner is actually a struggling reader and the teacher predicts any other category besides proficient reader, this would be equal to one. FP_{it} is the number of predictions where the teacher classifies the learner into the reference category, but the learner is not in the reference category. FN_{it} is the number of predictions where a teacher does not classify the learner into the reference category when in fact the learner falls into that category. Using Equations (1) to (4), we can then define the prediction accuracy rate as the proportion of predictions that were correctly classified:

$$Accuracy_{it} = \frac{(TP_{it} + TN_{it})}{(TP_{it} + TN_{it} + FP_{it} + FN_{it})} \quad (5)$$

The prediction accuracy of a classifier is a widely reported evaluation metric in the machine literature however it does not distinguish between how far away the misclassification is. For example, consider a group of 10 learners, 8 of whom are struggling readers the remaining 2 are proficient readers. Teacher A estimates that all learners are struggling readers and teacher B correctly classifies the eight to be struggling readers but classifies the two proficient readers as emerging readers. The two teachers would receive an identical accuracy score of 80% when the size of misestimation for Teacher B is actually smaller (one benchmark category) than that of Teacher A who underestimated the performance of the two proficient readers by two benchmark categories. Prediction accuracy would not distinguish between these two teachers, and we, therefore, need an additional measure to pick up this potential variation between teachers. The quadratic weighted kappa is commonly used to measure inter-rater agreement where agreement(accuracy) is weighted by the square of the deviation between the raters (Brenner & Kliebsch, 1996). It therefore overcomes the weaknesses of the prediction accuracy measure by accounting for the magnitude of the discrepancy between teacher estimates and measured learner performance.

We therefore have five different measures of teacher judgement accuracy, namely: prediction accuracy, rank correlation, point correlation, judgement error and quadratic weighted kappa. The next section specifies the model used to estimate the treatment effect on these outcome measures.

5.5. Identification Strategy

To identifying the causal impact of a program or intervention, we need to have a counterfactual of what would have happened in the absence of the intervention. We can express the causal impact for an individual as the difference in their outcome in the state where they receive the intervention (Y_i^1) and their outcome in the state where they do not receive the intervention (Y_i^0) (Cunningham, 2021):

$$\delta_i = Y_i^1 - Y_i^0 \quad (6)$$

We can also define the average treatment effect as the population mean of the individual treatment effect as well as two other population parameters, the average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATU) (Cunningham, 2021):

$$ATE = E[\delta_i] = E\left[Y_i^1\right] - E\left[Y_i^0\right] \quad (7)$$

$$ATT = E[\delta_i | D_i = 1] = E\left[Y_i^1 | D_i = 1\right] - E\left[Y_i^0 | D_i = 1\right] \quad (8)$$

$$ATU = E[\delta_i | D_i = 0] = E\left[Y_i^1 | D_i = 0\right] - E\left[Y_i^0 | D_i = 0\right] \quad (9)$$

Where D_i is a binary variable taking on the value of one if the individual receives the treatment and 0 if they do not. The difference in these parameters is that the ATE is the causal effect in the whole population while the ATT and ATU are the causal effects amongst the subpopulations of the treated and untreated respectively. The challenge of causal inference is that each individual has two potential outcomes, the one for the state of the world where they receive the intervention (Y_i^1) and one for the state of the world where they do not receive the

intervention (Y_i^0). In order to calculate the ATE, ATT and ATU, we need to observe both potential outcomes, but we can only observe one of these outcomes at a time. That is, a participant either receives the intervention or does not. This observed outcome can also be expressed as a switching function of the two potential outcomes (Cunningham, 2021):

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0 \tag{10}$$

Because we can only observe one potential outcome at a time, it is impossible to observe the causal impact for an individual and hence any of the population parameters. But we can still estimate the population parameters. Cunningham (2021) shows that the ATE can be decomposed as follows:

$$\begin{gathered} ATE = \pi ATT + (1-\pi) ATU \\ ATE = \pi\left(E\left[Y_i^1 | D_i = 1\right] - E\left[Y_i^0 | D_i = 1\right]\right) + (1-\pi)\left(E\left[Y_i^1 | D_i = 0\right] + E\left[Y_i^0 | D_i = 0\right]\right) \\ \therefore E\left[Y_i^1 | D_i = 1\right] - E\left[Y_i^0 | D_i = 0\right] \\ = ATE + \left(E\left[Y_i^0 | D_i = 1\right] - E\left[Y_i^0 | D_i = 0\right]\right) + (1-\pi)(ATT - ATU) \end{gathered} \tag{11}$$

Where π is the share of the population that receive the intervention. The second term on the right-hand side is the difference in the average outcomes in the absence of the intervention between those who received the intervention and those who did not. The third term is the difference in "returns" to the intervention between those who were treated and those who were untreated. The simple difference in mean outcomes between those who receive the intervention and those who do not receive the intervention is equal to the sum of the average treatment effect, selection bias and heterogeneous treatment effect bias (Cunningham, 2021).

The stable unit treatment value assumption (SUTVA) states that potential outcomes for a unit are unaffected by the treatments assigned to other units and that for each unit, there are no hidden variations in treatment which would lead to different potential outcomes (Imbens & Rubin, 2015). This means that there are no spillovers or externalities related to treatment assignment. The independence assumption states that treatments are assigned to individuals independent of their potential outcomes. Under SUTVA and the independence assumption in particular, the selection bias and heterogeneous treatment effect bias disappear, making the simple difference in mean outcomes a consistent estimator of the ATE:

$$\begin{gathered} E\left[Y_i^0 | D_i = 1\right] = E\left[Y_i^0 | D_i = 0\right] \because \left(Y_i^1, Y_i^0\right) \perp D \\ \therefore E\left[Y_i^0 | D_i = 1\right] - E\left[Y_i^0 | D_i = 0\right] = 0 \end{gathered}$$

(12)

$$E\left[Y_i^1|D_i=1\right]-E\left[Y_i^0|D_i=1\right]=\ E\left[Y_i^1|D_i=0\right]-E\left[Y_i^0|D_i=0\right]\ \because \left(Y_i^1,Y_i^0\right)\perp D$$
$$\therefore ATT-ATU=0$$
$$\therefore (1-\pi)(ATT-ATU)=0$$

(13)

By equations (12) and (13):

$$E\left[Y_i^1\middle|D_i=1\right]-\ E\left[Y_i^0\middle|D_i=0\right]=ATE$$

5.6. Non-compliance

One way to ensure that the independence assumption holds is to randomly allocate participants to receive the intervention or not. Treatment assignment would be exogenously determined and not depend on the potential outcomes of participants. However, except for very few cases, individuals are still free to decide to take up the treatment. Individuals may receive the intervention regardless of whether they were assigned to receive treatment or not. For example, assume we wanted to measure the impact of raising awareness of the COVID-19 pandemic and the benefits of vaccination on the likelihood that an individual gets vaccinated by randomly assigning households in a neighbourhood to attend a weekly workshop at the local community hall. Some households we assign to invite to the workshop may choose not to attend for whatever reason and we would not expect to see any major shifts in their outcomes.

Two issues may arise when estimating the impact of the workshop by calculating the simple difference in mean outcomes. The first is that the mean outcomes of households invited to attend the workshop would be brought down by the sub-group of households whose outcomes did not change because they did not actually attend the workshop. Holding all else constant, this may lead to underestimating the impact of the workshop on the likelihood of vaccination. On the other hand, households not formally invited to the workshop may still receive information informally through word of mouth and discussions with households that did attend. In this case, these households would essentially receive the treatment even without attending the workshop, impacting their likelihood of vaccination. Holding all else constant, the mean outcomes of those who were not invited to attend the workshop would be pulled up by this sub-group leading to an underestimation of impact. When we try to estimate the average treatment effect by taking the simple difference in mean outcomes between those who were invited to receive the intervention and those who were not, we would obtain a biased estimate.

The issue of non-compliance is common to programme evaluations especially when individuals are free to choose whether to opt in or out of the programme. The implication of non-compliance on causal inference is that although individuals' decision to select into program receipt may depend on their allocation to receive the program, allocation does not guarantee receipt. When estimating average treatment effects, the endogeneity of program receipt needs to be accounted for otherwise it will result in biased and hence internally invalid estimates. A common approach to addressing endogeneity in randomized control trials (RCT) is intent-to-treat (ITT) analysis. This is to estimate the effect of being randomly assigned to receive the treatment as opposed to the effect of the actual treatment itself. The random assignment process is exogenous and ensures that there is no difference in the average outcomes of those who are assigned to receive the treatment and those who are not. Continuing with the COVID-19 workshop example, we denote treatment allocation by the binary variable Z which takes the value one if a household was randomly invited to attend the workshop and 0 if not. The simple difference in mean outcomes can again be decomposed:

$$E\left[Y_i^1 \middle| Z_i = 1\right] - E\left[Y_i^0 \middle| Z_i = 0\right] = ATE + \left(E\left[Y_i^0 \middle| Z_i = 1\right] - E\left[Y_i^0 \middle| Z_i = 0\right]\right) + (1 - \pi)(ATT - ATU) \quad (14)$$

$$E\left[Y_i^0 \middle| Z_i = 1\right] = E\left[Y_i^0 \middle| Z_i = 0\right] \because \left(Y_i^1, Y_i^0\right) \perp Z$$
$$E\left[Y_i^1 \middle| Z_i = 1\right] - E\left[Y_i^0 \middle| Z_i = 0\right] = ATE \quad (15)$$

While ITT analysis will enable us to consistently estimate the average treatment effect using the simple difference in average outcomes, it fundamentally changes what the 'treatment' under consideration is. This treatment is the invitation to attend the workshop and not actual workshop attendance where information that may change outcomes is received by households. If the program has a positive relationship with the outcome, the average outcomes of the group who are invited to attend the workshop are diluted by those who did not attend the workshop. The ITT estimate underestimates the effect of attending the workshop, the treatment, even though it consistently estimates the effect of being invited to the workshop. It is, however, still helpful in thinking about potential program impact at scale where non-compliance is likely to occur because individuals are free to choose whether they take up a program that is freely available. Conditions at scale are much less controllable than

under an RCT, for example, and for this reason, the ITT may be the parameter of interest for audiences like policymakers and implementers.

A different approach to address endogeneity in program receipt is to use instrumental variable methods. Because program receipt status depends, at least partly, on allocation to receive the program, its contribution to variation in outcomes has two parts: the variation from receiving the program and the variation from being assigned to receive the program. In order to estimate the effect of program receipt on the outcome of interest, we need to isolate this effect from the effect of being assigned to receive the program. We can use assignment to receive the program as an instrument to isolate the effect of the actual program on the outcome of interest (Imbens & Angrist, 1994). We can express treatment status or program receipt (D_i) as a function of whether one is assigned to receive the treatment (Z_i):

$$D_i(Z) = D_i^0 + (D_i^1 - D_i^0)Z_i \quad (16)$$

Where D_i^0 is an individual's treatment status when Z_i is equal to 0 (when an individual is assigned to not receive the intervention) and D_i^1 is an individual's treatment status when Z_i is one (individual is assigned to receive the intervention). $D_i^1 - D_i^0$ is the causal effect of treatment assignment on treatment status and the causal effect of treatment assignment on the outcome $Y_i(Z, D)$ is:

$$\delta_i = Y_i\big(1, D_i(1)\big) - Y_i(0, D_i(0)) \quad (17)$$

Identifying the effect of treatment status, D_i, on the outcome, $Y_i(Z, D)$, is more complex because we need to account for how treatment status changes depending on assignment. The population can be partitioned into four groups of how individuals react to the instrument, treatment assignment (Angrist & Pischke, 2008):

1) **Compliers**: Individuals whose treatment status is positively determined by their treatment assignment ($D_i^1 = 1$ and $D_i^0 = 0$). For this group, if they are assigned to treatment, they will receive the treatment and if they are not assigned, they will not receive the treatment.
2) **Never-takers**: Individuals who never take-up treatment receipt regardless of whether they were assigned to receive it or not ($D_i^1 = D_i^0 = 0$)

3) **Always-takers**: Individuals who will always take-up treatment receipt regardless of whether they were assigned to receive it or not ($D_i^1 = D_i^0 = 1$)
4) **Defiers**: Individuals whose treatment status is negatively determined by their treatment assignment ($D_i^1 = 0$ and $D_i^0 = 1$). For this group, if they are assigned to treatment, they will not receive the treatment and if they are not assigned, they will receive the treatment.

In order for Z_i to be a valid instrument it needs to be correlated with treatment status and with the outcome measure. For never-takers and always-takers, their treatment status is unaffected by treatment assignment and therefore their treatment status is uncorrelated with the instrument. To overcome this, we would need to restrict the analysis to the population for whom their treatment status changes based on their assignment. This is the Local Average Treatment Effect (LATE) because we forgo how generalizable the estimate by estimating treatment effects amongst the select group of individuals whose treatment status is changed by their assignment (Imbens & Angrist, 1994). The LATE is commonly estimated using Two-Stage Least Squares, where the first stage is to estimate the effect of assignment on treatment status (Cunningham, 2021):

$$\gamma_i = D_i^1 - D_i^0 \tag{18}$$

By Equation (11), we can decompose this into:

$$\begin{aligned} &E\left[D_i^1 \middle| Z_i = 1\right] - E\left[D_i^0 \middle| Z_i = 0\right] \\ &= E\left[D_i^1 - D_i^0\right] + \left(E\left[D_i^0 \middle| Z_i = 1\right] - E\left[D_i^0 \middle| Z_i = 0\right]\right) + (1-\pi)(ATT_{Z,D} - ATU_{Z,D}) \end{aligned} \tag{19}$$

In order for the right-hand side of equation of equation (19) to reduce to γ_i, the SUTVA and independence assumptions need to hold. Under the LATE framework this is, the potential treatment status, D_i^1 and D_i^0, and the potential outcomes, Y_i^1 and Y_i^0 of each individual are uncorrelated with the potential treatment status of any other individual. Secondly, the potential treatment status and the potential outcomes are independent of the random assignment outcome such that $E\left[D_i^0 \middle| Z_i = 1\right] = E\left[D_i^0 \middle| Z_i = 0\right]$ and $E\left[D_i^1 \middle| Z_i = 1\right] = E\left[D_i^1 \middle| Z_i = 0\right]$. An implicit third assumption is that $E[D_i^1 - D_i^0]$ must actually exist which is only true when treatment assignment is correlated with treatment status. This means that individuals must base their

decision to take-up the intervention or not based on whether they were assigned to do so. Under these assumptions, equation (19) reduces to:

$$E\left[D_i^1 \middle| Z_i = 1\right] - E\left[D_i^0 \middle| Z_i = 0\right] = E\left[D_i^1 - D_i^0\right] \quad (20)$$

That is the difference in average treatment status for those who are assigned to receive the intervention and those who are assigned to not receive the intervention is the causal effect of treatment assignment on treatment status. For the complier population, this will be equal to one because $D_i^1 = 1$ and $D_i^0 = 0$. In order to derive the second stage which is the effect of treatment status on the potential outcomes, we need to make additional assumptions about the relationship between Z_i and D_i and their effect on Y_i. The exclusion restriction states that the effect of treatment assignment on the outcome $Y_i(D_i, Z_i)$ is only through treatment status such that:

$$Y_i(D_i, 1) = Y_i(D_i, 0) \quad (21)$$

The second assumption is monotonicity of the effect of Z_i on D_i. That is that the effect of treatment assignment on treatment status is either positive for all individuals or negative for all individuals:

$$\gamma_{1i} \geq 0 \; \forall \; i = 1, \dots, N \; or \; \gamma_{1i} \leq 0 \; \forall \; i = 1, \dots, N \quad (22)$$

The monotonicity assumption effectively excludes defiers whose treatment status is negatively determined by their assignment. The LATE is therefore only estimated amongst the sub-population of compliers. The second stage can be written as follows:

$$\begin{aligned} \delta_{LATE} &= \frac{\text{Effect of Z on Y}}{\text{Effect of Z on D}} \\ &= \frac{E\left[Y_i\left(D_i^1, 1\right) - Y_i\left(D_i^0, 0\right)\right]}{E\left[D_i^1 - D_i^0\right]} \\ &= E[(Y_i^1 - Y_i^0) | D_i^1 - D_i^0 = 1] \end{aligned} \quad (23)$$

By estimating the effect of Z through D on Y, we can in turn estimate the impact of the program for the complier sub-population directly and not just of the assignment. This estimation procedure relies on considerably more assumptions, in particular the exclusion restriction.

Continuing with the COVID-19 workshop example, the exclusion restriction means that being randomly invited to attend the workshop only affects an individual's propensity to get vaccinated based on whether they attend the workshop or not. There are many situations where this assumption could be violated. For example, if receiving the invitation to attend a workshop on COVID-19 induces participants to read up on the topic to decide the value of attending the workshop, the information they gather may change their orientation to vaccination and their propensity to get vaccinated even before they decide to attend or not attend the workshop. A violation of the exclusion restriction biases the IV estimates in a way that they cannot be interpreted as the LATE (Heckman, 1997). This assumption is also difficult to test because, for each individual, we would need information on their potential treatment status and outcomes in the two states: where they are assigned to receive the intervention and where they are assigned to not receive the intervention. We can only ever observe one of those outcomes at a time.

The LATE measures the treatment effect amongst the compliers. Compliers may be systematically different and distinct from other sub-groups in the population and the extent to which they are an interesting group from a policy perspective depends on the context. In a highly regulated sector where there is a culture of high compliance, the expected proportion of compliers at scale is likely to be significant and the LATE would be a parameter of interest. Depending on the program implementation, the RCT conditions may be more relaxed than if the program was rolled out by policymakers and hence the LATE may more closely mimic the expected program impact at scale. On the other hand, if the group of compliers is expected to be a trivial proportion and not representative at scale, the treatment effect amongst compliers would be uninformative to policymakers.

Given the number and nature of identifying assumptions that the LATE relies on, it is useful to explore another approach to addressing endogeneity in treatment receipt to estimate the direct effect of the intervention on outcomes. Because treatment receipt is non-random in the presence of non-compliance, we can apply strategies of causal inference from quasi-experimental studies. The most common of which is the difference-in-differences (DID) approach. This is to take the difference between the change in outcome over time (t) of the group that received the intervention ($D = 1$) and the change in outcome over time for the group that did not receive the intervention ($D = 0$):

$$E[\delta_i] = \left(E\left[Y_{D=1,i}|t=1\right] - E\left[Y_{D=1,i}|t=0\right]\right) - \left(E\left[Y_{D=0,i}|t=1\right] - E\left[Y_{D=0,i}|t=0\right]\right)$$

(24)

Where t takes the value 0 for before the intervention and 1 for after the intervention is introduced. From equation (10) and the fact that we can only observe one potential outcome for each individual, we can denote $Y^1_{D=1,i}$ to be the outcome when $D=1$ and the intervention has been introduced, $Y^0_{D=1,i}$ to be the outcome when $D=1$ but the intervention is yet to be received and $Y^0_{D=0,i}$ to be the outcome when $D=0$ regardless of time:

$$\delta_{DID} = \left(E\left[Y^1_{D=1,i}|t=1\right] - E\left[Y^0_{D=1,i}|t=0\right]\right) - \left(E\left[Y^0_{D=0,i}|t=1\right] - E\left[Y^0_{D=0,i}|t=0\right]\right)$$

(25)

$$\begin{aligned}
&= \left(E\left[Y^1_{D=1,i}|t=1\right] - E\left[Y^0_{D=1,i}|t=0\right]\right) - \left(E\left[Y^0_{D=0,i}|t=1\right] - E\left[Y^0_{D=0,i}|t=0\right]\right) + \boldsymbol{E\left[Y^0_{D=1,i}|t=1\right]} \\
&\quad - \boldsymbol{E\left[Y^0_{D=1,i}|t=1\right]} \\
&= \left(E\left[Y^1_{D=1,i}|t=1\right] - E\left[Y^0_{D=1,i}|t=1\right]\right) + \left(E\left[Y^0_{D=1,i}|t=1\right] - E\left[Y^0_{D=1,i}|t=0\right]\right) \\
&\quad - \left(E\left[Y^0_{D=0,i}|t=1\right] - E\left[Y^0_{D=0,i}|t=0\right]\right) \\
&\therefore E[\delta_i] = ATT + \left(E\left[Y^0_{D=1,i}|t=1\right] - E\left[Y^0_{D=1,i}|t=0\right]\right) - \left(E\left[Y^0_{D=0,i}|t=1\right] - E\left[Y^0_{D=0,i}|t=0\right]\right)
\end{aligned}$$

(26)

The DiD approach will consistently estimate the average treatment effect on the treated when $\left(E\left[Y^0_{D=1,i}|t=1\right] - E\left[Y^0_{D=1,i}|t=0\right]\right) - \left(E\left[Y^0_{D=0,i}|t=1\right] - E\left[Y^0_{D=0,i}|t=0\right]\right)$ is equal to zero. This is when the change in outcome over time that would have experienced by those receiving the intervention, in the absence of the intervention is equal to the change in outcomes over time experienced by those who do not receive the intervention. This is the parallel trends assumption and under it, equation (26) simplifies to:

$$\delta_{DID} = ATT$$

(27)

The average treatment effect amongst the treated estimates the actual effect of receiving the program. In contexts where policymakers cannot force participation and where costs increase as a result of participation, it is a useful parameter to estimate. This could be for example a cash transfer, training, or resource provision program. To make decisions about whether such a program should be implemented at scale, policymakers need information on what the impact

of the program will be for people who opt in. Even if costs are not driven up by program participation, it is still useful to estimate the impact of receiving the program in order to make decisions around the value of mobilising resources to encourage individuals to opt into the program. Estimating the ATT using DiD relies on the parallel trends assumption which is an exclusion restriction that cannot be tested. Even if one has access to data from multiple periods prior to the introduction of the program that identifies parallel trends, one cannot conclude that these trends would have continued in the absence of the intervention. We are still able to look for evidence to support the parallel trends assumption even if we are unable to conclusively say that it holds. Although DiD does allow for imbalance in the pre-treatment level of the outcome of interest and other covariates between the treatment group and the counterfactual, the more similar the two groups are pre-treatment, the more plausible it is that the parallel trends assumption holds. Randomizing treatment assignment, even with imperfect compliance, is one way to reduce the differences between the two groups prior to the introduction of the program.

To summarize, the simple difference in average outcomes of those who receive the intervention and those who do not is a consistent estimator of the ATE under the SUVTA and independence assumptions. Randomisation is one way to ensure that the independence assumption holds. However, under non-compliance, random assignment to treatment does not imply program receipt. We can use programme assignment as per the outcome of randomisations and estimate the ITT effect. This is the effect of treatment assignment on the outcome of interest. While this may be a policy-relevant parameter, it is not informative on the direct impact of the treatment on outcomes. To estimate the actual treatment effect, we can use instrumental variable methods to estimate the local average treatment effect amongst the sub-population who comply with their treatment assignment. This approach relies on several assumptions, which are difficult to test. An alternative strategy is to use the DID approach which will yield a consistent estimate of the average treatment on the treated, provided that the parallel trends assumption holds. Therefore, to estimate the direct impact of treatment, both strategies require us to forgo the external validity of the estimate because we are estimating amongst a select group of individuals although the ATT suffers from less selection than the LATE. In the next sub-section, we discuss the context of the present study and formally state our estimation strategy.

5.7. The present study

The first research question relates to measures of teacher judgement accuracy as a proxy for teacher knowledge of learning levels. The outcome measures have been discussed in Section 5.4 and these are the Kendall-Tau Rank Correlation, Pearson Correlation, Average Judgement Error and Prediction Accuracy. We will investigate this question through a descriptive analysis of these measures. The second research question is to investigate the effect of the intervention on teacher knowledge of learning levels. We define our regression model as:

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \quad (28)$$

Where:

$$\therefore \beta_0 = E\left[Y_i^1 \middle| D_i = 0\right]$$
$$\therefore \beta_1 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

The coefficient of interest is β_1 which under the SUTVA and Independence assumptions is a valid estimate of the average treatment effect. Assignment to treatment was randomised within school and grade but teachers still needed to attend training in order to receive the intervention. We can reasonably expect a degree of non-compliance which means assignment does not imply treatment receipt. Our first-order interest is in estimating the effect of intervention receipt because this is a pilot study of how the newly established benchmarks in African languages can be used in the classroom. In order to recover the treatment effect, our preferred strategy is to use the DiD approach to estimate the average treatment effect on the treatment:

$$y_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 (D * T)_{it} + \alpha_i + \varepsilon_{it} \quad (29)$$

Where y_{it} is the outcome of interest (judgement error, rank correlation, point correlation, prediction accuracy and quadratic weighted kappa) of teacher i in time t, D_i is the treatment status of teacher with 0 indicating control and 1 indicating treatment receipt, T_t is a binary variable indicating time, defined as 0 for baseline and 1 for endline, $(D * T)_{it}$ is an interaction term between treatment status and time, α_i is the time-invariant error term for each teacher or the individual fixed effect and ε_{it} is the time-varying random error term for each teacher in time

t. Under this specification, the coefficient of interest is β_3, also referred to as the difference-in-difference term:

$$\beta_3 = \left(E\left[Y_{D=1,i}|T=1\right] - E\left[Y_{D=1,i}\,|t=0\right]\right) - \left(E\left[Y_{D=0,i}|t=1\right] - E\left[Y_{D=0,i}|t=0\right]\right) \quad (30)$$

This will produce a consistent estimate of the ATT if the parallel trends assumption holds. We will test the level difference in means in the period before the intervention was introduced to test if the parallel trends assumption is likely to hold. We also investigate whether there is balance between groups on a range of baseline covariates. As noted by Kahn-Lang and Lang (2019), a level difference that is statistically insignificant from zero in the period before the intervention is introduced is neither a necessary nor sufficient condition for the parallel trends assumption to hold. However, given that the parallel trends assumption cannot be empirically tested, the level difference test will be used to motivate our view on whether it is likely to hold or be violated. If the groups are inherently similar on the level difference in means, it makes the parallel trends assumption more plausible. We also use equation (28) and substitute in the binary random assignment variable Z_i in place of D_i to estimate the ITT as a conservative estimate of the ATE at scale. Finally, we calculate the LATE as the optimistic estimate of intervention effects for compliers. These three parameters together represent the full range of intervention effects which will be an important input of whether this intervention can be taken forward.

5.7.1. Statistical Power Analysis

Our estimation strategy is such that we try and estimate the true effect of the intervention using the data we observe in the study. Implicitly, we are testing the null hypothesis, H_0, that the intervention had no impact on the outcome, teacher knowledge of learning levels. The significance level of the hypothesis test is the probability that we falsely reject the null hypothesis and conclude that the intervention had an effect when it did not. Whereas the power of the test is the probability that we do not fail to reject the null hypothesis when the null hypothesis is false, i.e., we conclude that the intervention had no effect when in fact it did. When we estimate the effect of the program, we want to maximize the power while minimizing the significance level such that our estimated program effect is as close as possible to the true program effect. We can investigate whether this is the case by calculating the minimum detectable effect size for a given sample size and make judgements on whether this is realistic based on how we expect the intervention to affect teacher outcomes. This is done by computing the following equation as specified by Duflo, Glennester and Kremer (2007):

$$MDE = \left(t_{1-K} + t_{\alpha/2}\right) \times \sqrt{\frac{1}{P(1-P)}} \times \sqrt{\frac{\sigma^2}{N}}$$

Where $1-K$ is the statistical power, conventionally set at 80%, α is the significance level conventionally set at 5%, N is the total sample size, P is the proportion of the sample in the treatment group and σ^2 is the variance of the outcome. We estimated the minimum detectable effect sizes based on the baseline distributions of the outcome measures. These are reported in Table 8.

Table 8: Minimum Detectable Effect Sizes

	(1)	(2)	(3)
Variable	Mean	SD	MDE
Kendall-Tau Correlation	0.551	0.244	0.116
Pearson Correlation	0.665	0.289	0.137
Average Judgement Error	11.450	6.761	3.181
Prediction Accuracy	0.740	0.079	0.037
Quadratic Weighted Kappa	0.457	0.279	0.131

Note: Table reports minimum detectable effect sizes determined through statistical power analysis across outcome measures. We set the significance level to 5% and the Power to 80%. The number of teachers in the untrained group is 125 and therefore the ratio of treatment to control is 0.64. MDE is measured in the units of each outcome measure

The study is powered to detect differences in the Kendall-Tau Correlation of 0.116 or larger. This means that the study will be able to detect an effect at the 5% level of significance if the difference in average correlation for treatment relative to control at endline is 0.116 or more. Any difference in correlation below this value will be statistically insignificant.

6. Sample Description

6.1. Learner Performance

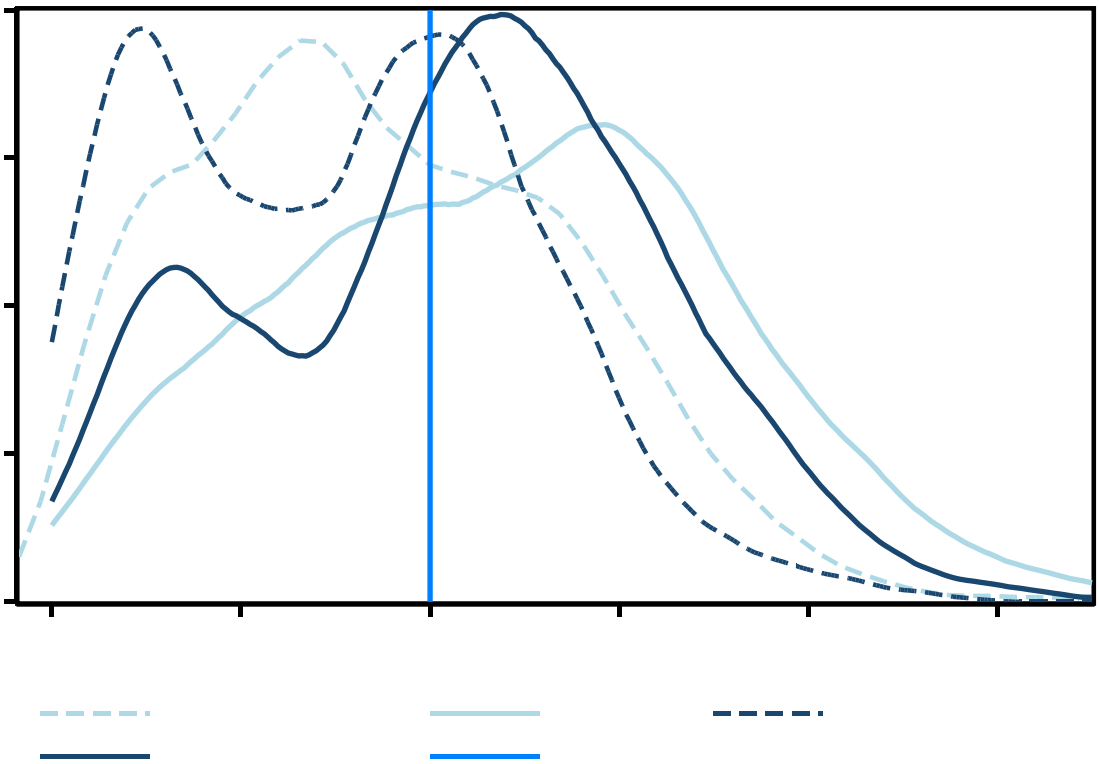
Our theory of change (Figure 2) outlines improved learner performance as the main outcome of differentiated instruction. Due to the modest scope and scale of this pilot study and the fact that teachers had six months with the intervention, we do not expect to see any significant shifts in pedagogic practices and hence in learning outcomes. The main outcome of the study is teacher knowledge of learner levels which in the context of differentiated instruction is an

important program output. We measure learner performance in order to have an objective comparison with teacher estimations. We will however report learner performance for informational purposes and to understand where learners are in relation to the newly established reading benchmarks. This section summarizes the results of the independently conducted EGRA assessments at both baseline and endline. Learners who could not sound five letters correctly in their home language did not attempt the oral reading fluency task, this was an opt-out rule implemented to reduce stress to learners who could not read.

6.1.1. Letter Sound Knowledge

Learners were asked to sound out as many letters in their home language from the provided chart within one minute. Figure 6 gives the distribution of learner letter sound knowledge. The benchmark of 40 correct letters per minute (CLPM) is the same for Grade 2 and 3 learners as this is the level they should have reached by the end of Grade 1. For Sesotho-Setswana learners, the distribution at baseline was bimodal around 10 and 40 CLPM. By endline, the distribution had shifted rightward and unimodal around 55 CLPM. For Nguni learners, the baseline distribution of letter sound knowledge was skewed to the right with a peak around 25 CLPM. By endline, this distribution was left-skewed with a peak around 60 CLPM although the variance is larger than for Sesotho-Setswana learners. As expected, both groups of learners improved in their letter sound performance over the period and the proportion of learners meeting the letter sound benchmark increased overall.

Figure 6: Letter Sound Knowledge Benchmark Categories



6.1.2. Oral Reading Fluency

In this task, learners were presented with a passage and given one initial minute to read the passage and an additional two minutes to finish the passage for the comprehension sub-task. The measure of oral reading fluency is how many words the learner reads correctly in the first minute. Learners who read the first five words incorrectly had their assessment end early and are classified as Non-Readers. Figure 7 shows the distribution of Correct Words per Minute for Grade 2 learners. The distribution of CWPM for Nguni learners is skewed to the left and the proportion of density to the right of the 20 CWPM benchmark is relatively small. The majority of learners were not meeting the Grade 2 benchmark. By endline, the distribution has shifted right and increased in variance but the density around non-readers (0 CWPM) was relatively unchanged and the peak was still around 8 CWPM. The majority of learners were still not meeting the benchmark but there was an increase in the proportion of learners who were.

For Sesotho-Setswana learners at baseline, the density around non-readers was much larger and very few learners were meeting the benchmark of 40 CWPM. By endline, the proportion

of non-readers had decreased significantly, and more learners were meeting the benchmark although not many more. Learners' oral reading fluency therefore improved for both groups but at the top end of the distribution for Nguni and mostly at the bottom end for Sesotho-Setswana. The average Grade 2 learner was reading 12 words correctly per minute in isiXhosa, 8 in isiZulu, 9 in Setswana and 11 in Sepedi at baseline (Table 9). This increased to 18 in isiXhosa, 13 in isiZulu, 19 in Setswana and 21 in Sepedi at the endline. The average Grade 2 learner in all provinces does not meet the reading benchmark although this is only by two words in isiXhosa at endline. The average learner would also be classified as a struggling reader in all provinces except the Eastern Cape where this learner is an emerging reader.

Figure 7: Oral Reading Fluency Benchmark Distribution - Grade 2

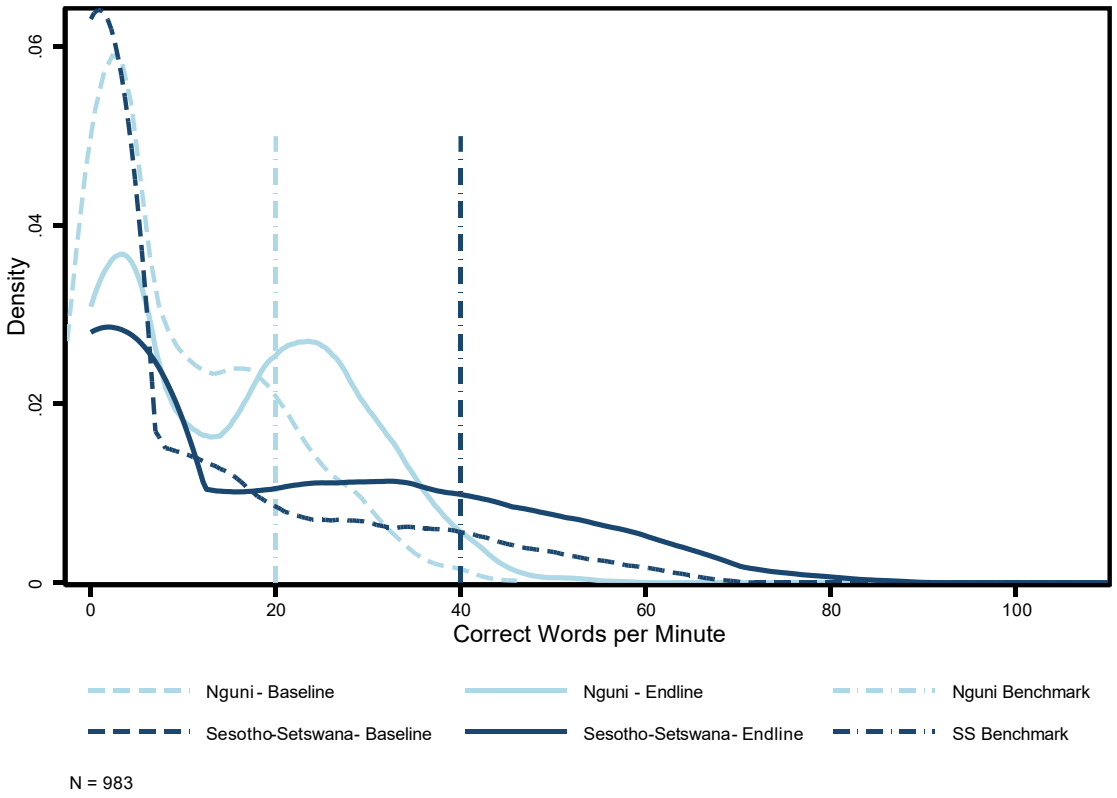


Table 9: Average CWPM - Grade 2

	Baseline		Endline	
	Mean	Std Dev	Mean	Std Dev
Eastern Cape	12,08	10,57	18,33	11,96
Limpopo	11,38	16,16	20,40	20,99
Mpumalanga	8,29	8,65	12,92	12,40
North-West	9,18	14,61	18,27	20,96

Note: Table reports the average correct words read per minute for Grade 2 learners at baseline and endline by province. N = 983

For Grade 3 learners across both language groups, we see an almost identical distribution in Figure 8 as for Grade 2 learners in Figure 7. For Nguni learners, more learners are moving from being struggling to emerging readers with a stable non-reader proportion. For Sesotho-setswana, learners are moving from non-readers to struggling readers and from struggling to emerging, but few are moving from emerging to proficient. The result is an increase in the proportion of struggling readers, a decrease in the proportion of non-readers and a small increase in the proportion of proficient readers. At endline, the average grade 3 learner was correctly reading 23 words in isiXhosa, 18 in isiZulu, 26 in Setswana and 32 in Sepedi (Table 10). This means that by endline the average grade 3 learner does not meet the benchmark and would be classified as a struggling reader in all provinces.

Figure 8: Oral Reading Fluency Benchmark Distribution - Grade 3

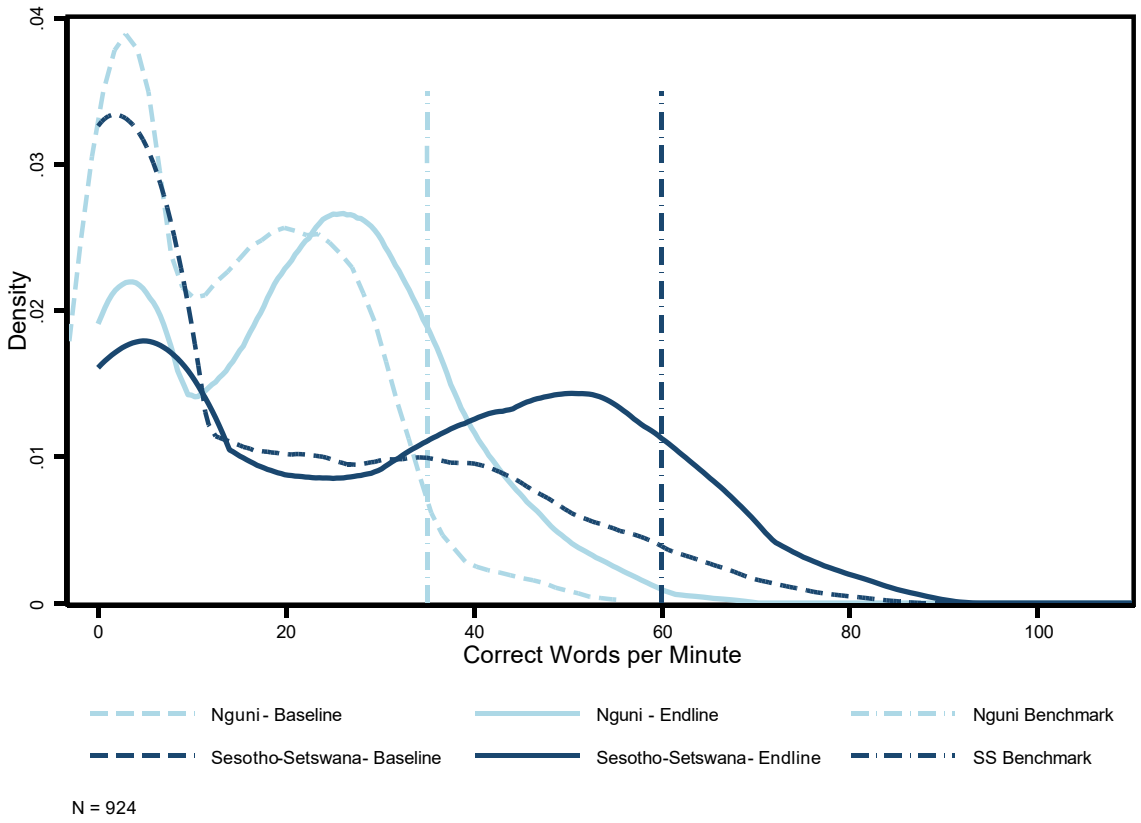


Table 10: Average CWPM - Grade 3

	Baseline		Endline	
	Mean	Std Dev	Mean	Std Dev
Eastern Cape	15,69	11,48	24,65	13,82
Limpopo	17,29	19,58	33,00	23,64
Mpumalanga	12,39	11,36	18,32	14,48
North-West	16,67	20,36	27,38	23,78

Note: Table reports the average correct words read per minute for Grade 2 learners at baseline and endline by province. N = 924

Although the sample is not representative, the reading levels are similar to that found in provincially representative and other larger localised studies (Menendez & Ardington, 2018). The better performance for Sesotho-Setswana home language learners is in part due to the difference in average word length between the language families. This is reflected by the different word length benchmark levels within the same grade across the two language families. For instance, the benchmark for Grade 2 Nguni language learners is 20 CWPM while that of Grade 2 Sesotho-Setswana learners is 40 CWPM. In addition, the Progress in International Reading Literacy Study (PIRLS) 2016 found that Sepedi and Setswana Home Language Learners had the highest proportion of learners who did not meet the lowest international benchmark (Howie et al., 2017).

The learner performance in our sample highlights two important issues: There is a spread of learners between benchmark categories which means that learning levels are heterogeneous, and the majority of learners do not meet the reading benchmarks. The implication is that learners are not on track to being able to read fluently in their home languages. Effective teaching strategies to bring a learner who cannot read a single word correctly (e.g. non-reader) closer to fluency are very different from a learner who reads slowly (e.g. emerging reader). This is because the emerging reader already has decoding skills that need to be strengthened through for example developing word recognition, but a non-reader needs to still build those skills through building phonological awareness (Wilsenach, 2019). With such heterogeneity in learning levels and hence heterogeneity in learning needs, differentiated instruction becomes important in improving learner performance across the achievement distribution.

6.2. Teacher Characteristics

The average characteristics of the 197 teachers who were interviewed at both baseline and endline are shown in Table 11 for the full sample and separately for trained and untrained teachers. The average teacher has 17 years of experience, 14 of which are in teaching in the foundation phase. Thirty-five (35) percent of teachers have a bachelor's degree and 68% are formally trained to teach in the foundation phase. The minimum class size is 27 with a maximum of 70 and a mean of 46. In the TIMMS 2019 analysed by Spaull et al. (2022), the average class size was also 46 and they found that not only did class size increase between 2015 and 2019, but the provinces with the largest average class sizes were Kwa-Zulu Natal, Limpopo and Mpumalanga. Two of the provinces with the largest class sizes are study sites for the present study.

Table 11: Average Teacher Characteristics

Variable	All		Trained		Untrained		P-Value
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	
Experience (Years)	17,21	10,75	16,84	10,34	17,86	11,47	0,52
Experience in Foundation Phase	13,76	10,30	13,86	9,98	13,57	10,90	0,85
Has a Diploma	0,30	0,46	0,30	0,46	0,32	0,47	0,73
Has a Degree	0,35	0,48	0,39	0,49	0,28	0,45	0,11
Training in Foundation Phase	0,68	0,47	0,70	0,46	0,63	0,49	0,26
Previously EGRA Trained	0,52	0,50	0,51	0,50	0,53	0,50	0,79
Class Size	45,92	7,76	46,05	7,59	45,71	8,11	0,77
Groups Learners	0,76	0,43	0,74	0,44	0,78	0,42	0,60
Percent Correctly Classified	0,35	0,20	0,35	0,20	0,35	0,19	0,81
Spearman Rank Correlation	0,63	0,27	0,63	0,28	0,65	0,26	0,59
Kendall-Tau Rank Correlation	0,55	0,24	0,54	0,25	0,57	0,24	0,39
Average Learner CWPM	12,86	7,24	12,82	7,34	12,93	7,09	0,92
Average Learner CLPM	32,93	11,26	33,06	11,27	32,69	11,33	0,83
Number of Teachers	197		125		72		

Note: SD = Standard Deviation. P-Value of the significance in difference between the trained and untrained groups. A value of 0.05 or less indicates that the difference in the mean values of the groups is statistically significant or different from zero at the 5% significance level. Sample excludes the school where there were no trained teachers.

The final column reports the p-values from a two-sample t-test of differences in the means between two groups from the same distribution. There are no statistically significant differences at the 10% significance level in the average professional characteristics, outcome measures and learner performance measures between the treatment and control. In other words, the treatment and control were not systematically different in their characteristics which

are correlated with outcomes and in the outcome measures. This is true even when we account for attrition. The balance between the two groups supports the plausibility of the parallel trends assumption.

6.3. Home Language Teaching Practice

The CAPS Curriculum guides teachers with prescriptions on the minimum instructional time to be spent on certain teaching activities. Table 12 summarises these prescribed frequencies and Figure 9 plots the frequencies at which teachers report teaching the various activities at baseline separately for grade 2 and 3 teachers.

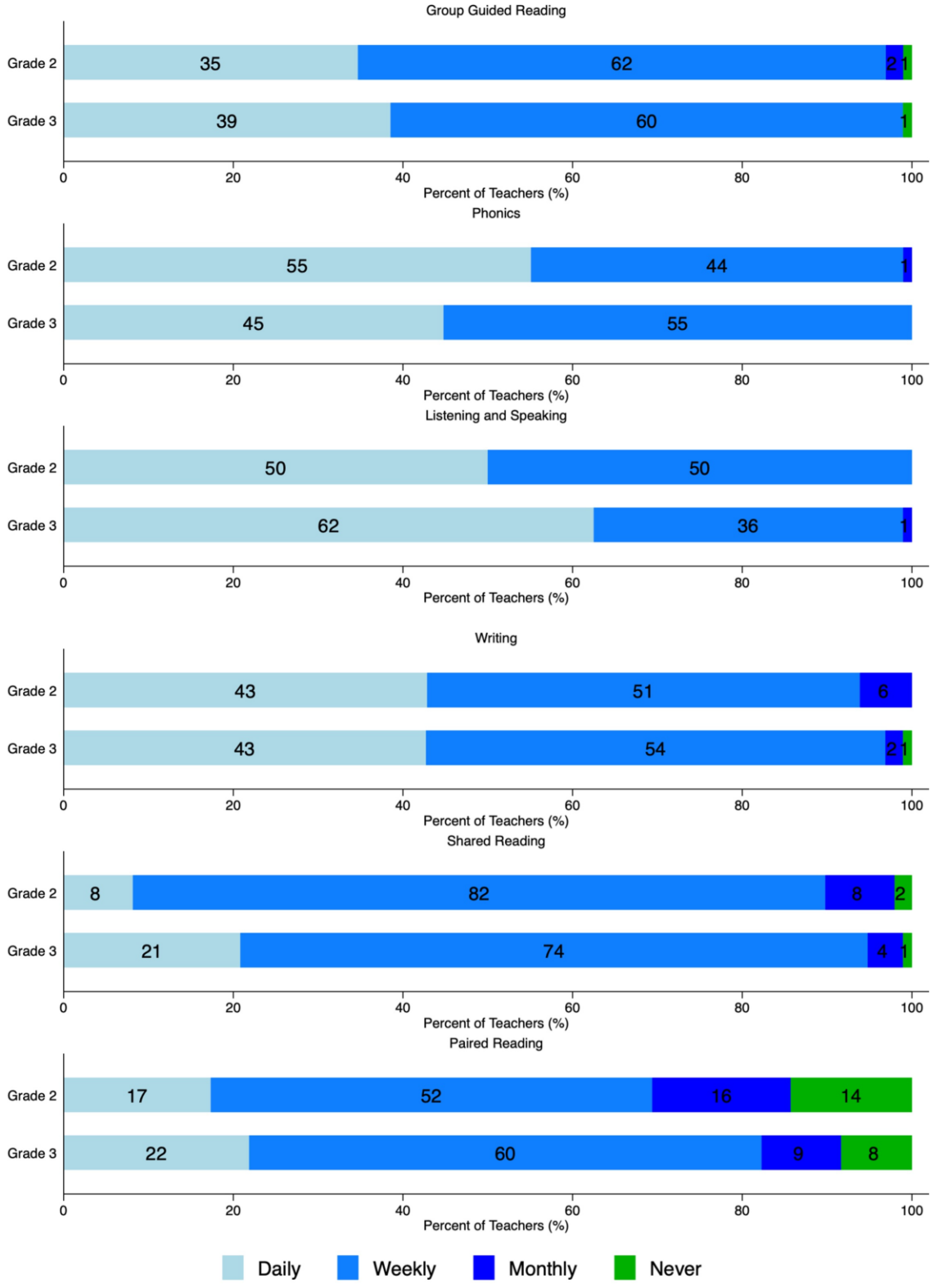
Table 12: Prescribed Frequencies of Home Language Activities

Activity	Grade 2	Grade 3
Group-Guided Reading	Daily	Daily
Phonics	Daily	2-4 times a week
Listening & speaking	2-4 times a week	2-4 times a week
Writing	2-4 times a week	2-4 times a week
Shared Reading	2-4 times a week	2-4 times a week

Note: Table reports the frequency at which various home language activities are supposed to be taught as per CAPS Curriculum. Source: (DBE, 2011a)

Group-guided reading requires that teachers know the learning levels of learners in order to put them into ability groups for differentiated instruction. The CAPS curriculum requires that teachers conduct group guided reading daily but the majority of teachers report teaching it less often. For all other home language activities, the majority of teachers report conducting those activities at least as often as prescribed. The finding that teachers teach group guided reading less than required is consistent with the findings of Cilliers et al. (2018).

Figure 9: Frequency of Home Language Teaching Activities



6.4. EGRA Exposure and Use

Table 13 reports teachers' exposure to EGRA where exposure is defined as having either received training directly from the DBE or having previously conducted the assessment. The proportion of teachers who reported some exposure to EGRA overall was 57%. This is in line with our expectations given that just under half of the schools had received EGRA according to the DBE database. Surprisingly, Mpumalanga had the lowest proportion of teachers with exposure to EGRA despite the fact that nine of the 10 schools in Mpumalanga were part of the DBE EGRA roll-out. On the other hand, 54% and 70% of teachers in North-West and Limpopo respectively reported being exposed to EGRA even though their schools were not part of the DBE EGRA roll-out.

Table 13: Teacher prior EGRA exposure by Province

Variable	North West	Limpopo	Mpumalanga	Eastern Cape	Overall
EGRA Trained	39%	65%	36%	54%	48%
EGRA Experienced - No Training	15%	5%	2%	13%	9%
No EGRA Exposure	46%	30%	62%	33%	43%
Num. of Teachers	59	54	58	54	**225**

Note: The sample excludes three teachers who opted out of the interview before reaching this question.

Baseline data were collected in May which corresponds to roughly the middle of the second school term. According to the DBE EGRA programme, Grade 2 and 3 teachers are required to have completed at least one assessment by this point. Of the 128 teachers who have had exposure to EGRA, 74% had completed at least one assessment round by halfway through term 2 (Table 14). This proportion differs by province from as low as 38% in Mpumalanga to 76% in the Eastern Cape.

Table 14: Proportion of teachers with exposure to EGRA administering EGRA by May

Province	Completed EGRA by May
Eastern Cape	76%
North-West	69%
Mpumalanga	38%
Limpopo	74%
Overall	74%

Note: Table reports the proportion of teachers who had EGRA exposure through experience or formal training (N = 128) who completed at least one assessment by May.

6.5. EGRA Assessment Practice

We draw on baseline teacher interviews to generate insights into home language formative assessment practices with a particular focus on the use of EGRA. We restrict the sample to teachers who had completed at least one assessment by May. If a teacher had not completed the EGRA by this point, their responses to questions regarding their EGRA assessment and reporting practices may be influenced by recall bias. This means that teachers with less reliable memory might provide responses that don't accurately represent their past actions. To mitigate the potential recall bias, we excluded those who had not conducted an assessment by the time of baseline data collection. Since the percentage of teachers with exposure to EGRA was lower than expected, we modified the teacher questionnaire to focus on general formative assessment practices at the endline. Therefore, our investigation of teachers' formative assessment practices using the EGRA tool primarily relies on baseline data. We specifically analyse the assessment practices related to EGRA and the use of assessment results for the 95 teachers who had completed at least one EGRA at the baseline.

Figure 10 shows the distribution of the average length of an EGRA with a single learner for teachers who had completed at least one assessment by May. Roughly 35% of teachers report that on average each assessment takes up to five minutes per learner. A further 33% say each assessment takes between five and 10 minutes. The average duration of EGRA is 8.74 minutes but the median duration is five minutes. This is because of the teachers who report taking up to 60 minutes per learner artificially increase the mean. Given the skewed distribution, the median is the more appropriate summary statistic of the distribution of average EGRA duration. Considering the distribution of class size depicted in Figure 11, the average class size is around 45 learners. With a median EGRA duration of five minutes per learner, this implies that it would take a teacher roughly four contact hours to assess the whole class.

Figure 10: Distribution of Average EGRA duration

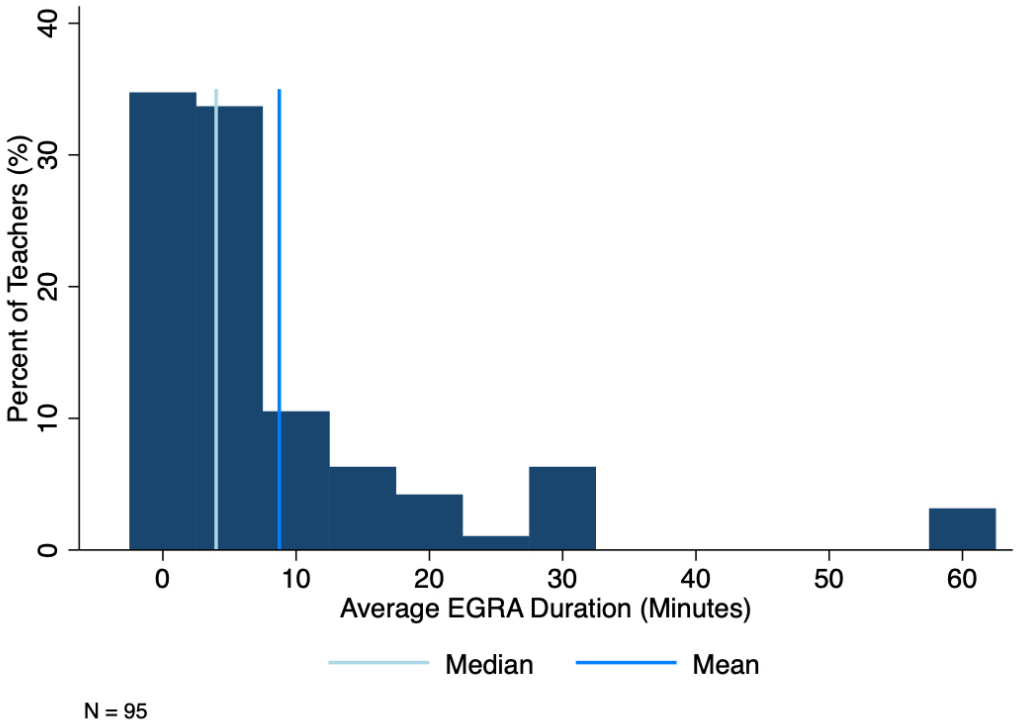
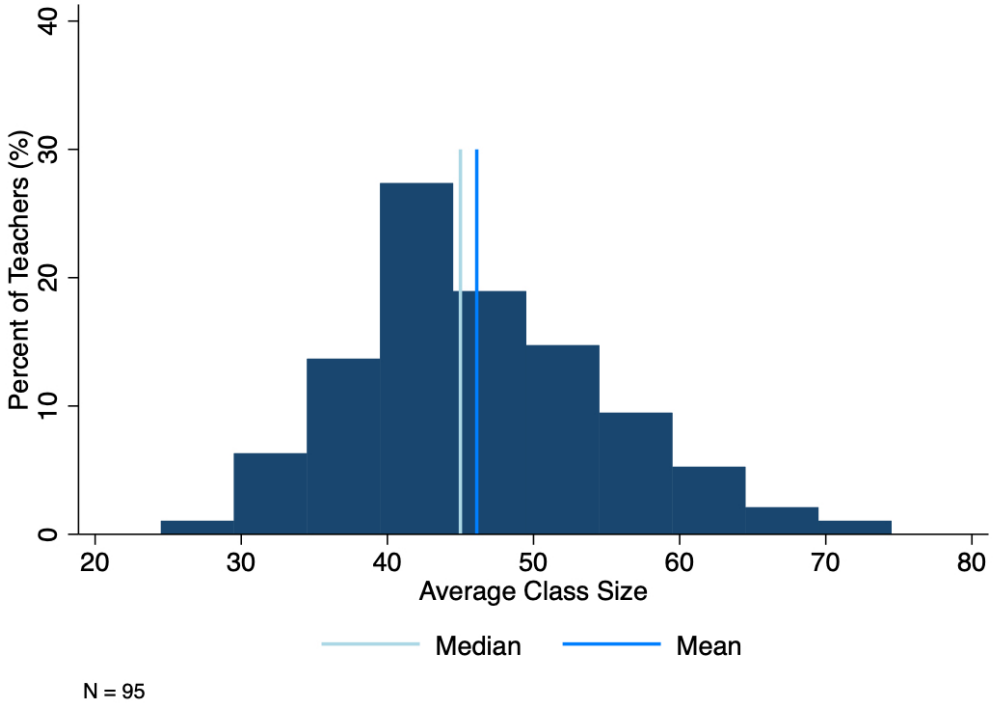


Figure 11: Distribution of Class Size



According to the CAPS curriculum, Home Language instruction is allocated seven hours per week (Table 12). For teachers to administer a single round of home language assessment using the EGRA, they would have to forgo almost an entire week of home language instruction. Some teachers have between 60 and 70 learners in their classes, and for these teachers, it would take between five and six contact hours to assess the whole class. Accounting for the time it takes to finish an assessment with one learner and to start it with the next, these teachers would either need to use instructional time allocated to other subjects or take over a week to assess the whole class.

The “lost” instructional time can be mitigated through the use of effective classroom management strategies during assessment time. For example, a teacher may assign individual work to learners to work silently while a single learner is assessed. Such strategies would ensure that learning time is not lost even if a teacher is not providing whole-group instruction. The various strategies employed by teachers while they complete an assessment with an individual learner are summarized in Table 15. The majority of teachers (62%) have their learners work in their DBE workbooks or some other worksheet. Twenty percent (20%) of teachers facilitate a reading corner and a further 13% have learners complete a writing activity. Although the majority of teachers still keep learners engaged in learning activities, 5% of teachers do not. For these teachers, the structured learning time would effectively be lost during a week of home language assessment.

Teachers were asked how they believe EGRA administration could be made easier for them as well as what they use the results of EGRA for once assessment is completed. For both these questions, teachers could have selected more than one potential response. In terms of making EGRA easier to administer, 33% of teachers responded that decreasing the time burden of assessment would make it easier. As mentioned above, conducting a single round of assessment on average would take a week of home language instruction time and teachers are required to conduct a minimum of three assessment rounds in the year in Grades 2 and 3. With the pressure of curriculum coverage and learners who are performing below their grade level, it can be difficult to conduct three rounds of assessment and effectively lose three weeks of home language instruction time. This may potentially be a contributing factor to 26% of teachers who have been exposed to EGRA not having conducted at least one assessment by half through the second school term.

Twenty-three percent (23%) and 21% of teachers respectively, also believe that more training and resources would make it easier for them to administer the EGRA. It is noteworthy that this is amongst teachers who have both been exposed to EGRA through training or conducting

one themselves and have completed at least one assessment round by halfway through the second term. Thirty-five percent (%) of teachers gave other responses for how to make EGRA administration easier (Other – Table 15). The majority of these teachers spoke about increasing the time given for each learner to read and making the levelled assessment passages easier for learners to understand. These teachers lack understanding of the EGRA tool, what it measures and how. This may speak to how teachers are trained on the EGRA tool. With regards to how teachers use EGRA, the biggest proportion of teachers use it for identifying learners for remediation or consolidation (Table 16).

Table 15: Teacher EGRA Practices

Variable	**Mean**
Average EGRA Time per Learner (Minutes)	8,74
Classroom management during assessment:	
Workbooks/worksheets	62%
Reading	20%
Free time	5%
Writing	13%
Avenues for Easier Administration:	
Decrease Time Burden	33%
More Training	23%
More Resources	21%
Other (None of the above)	35%
Num. of Teachers	95

Note: Sample includes only teachers who had prior exposure to EGRA and who had completed at least one assessment by May

Table 16: Teacher use of EGRA Results

Variable	Mean
Report to DBE	13%
Create Group Guided Reading Groups	23%
Identify Learners for Remediation/Consolidation	44%
Differentiate Instruction	11%
Set Lesson Plans	17%
Classroom Management (excl. GGR)	11%
Assign Materials to Learners	13%
Report to School Management	9%
Other (None of the above)	3%
Num. of Teachers	95

Note: Sample includes only teachers who had prior exposure to EGRA and who had completed at least one assessment by May

In summary, we find that many teachers have no exposure to EGRA and this does not align with the DBE rollout. While it is possible that the teachers are new, it is more likely that it reflects absenteeism or that training was too light-touch. This is supported by the number of teachers reporting the need for further training. Teachers also manage the classroom by giving other learners individual work like their workbooks, worksheets or reading when they conduct individual assessments. Therefore, even though it takes long to assess the whole class, according to teacher self-reports, learners are not losing out on learning during that time.

7. Results

7.1. Baseline Assessment Efficacy

To answer the first research question which is to measure teacher judgement accuracy, we examine teachers' knowledge of the reading proficiency levels of learners in their class at baseline. We compare teacher rankings and oral reading fluency estimates against the results from the EGRA conducted by our field teams at baseline. Teachers were shown the same passage that was used in our learner assessment and asked to estimate how far each of the

10 learners that we assessed could read within a minute. Figure 12 and Figure 13 plot the teachers' estimate of learners' correct words read per minute against our measured correct words read per minute for each of the 10 sampled learners per teacher. Observations that lie on the line of equality indicate that teachers correctly estimate their learners' performance. The majority of the points lie above this line which means that teachers tended to overestimate the performance of their learners at baseline.

We also find that teachers tend to do a poor job of estimating reading proficiency at all points along the achievement distribution. By way of illustration using Figure 12, for learners who are reading no words correctly and those who are reading 20 words correctly, the Grade 2 Nguni Benchmark, there is a full range of teacher responses. This suggests that it doesn't matter whether a learner is more proficient or not, teachers still misestimate their performance. Generally, it seems that there is a weak relationship between teacher estimates and learner performance. To explore this, we calculate the strength of the association between the two measures.

Figure 12: Teacher Estimate vs Learner Performance at Baseline – Nguni languages

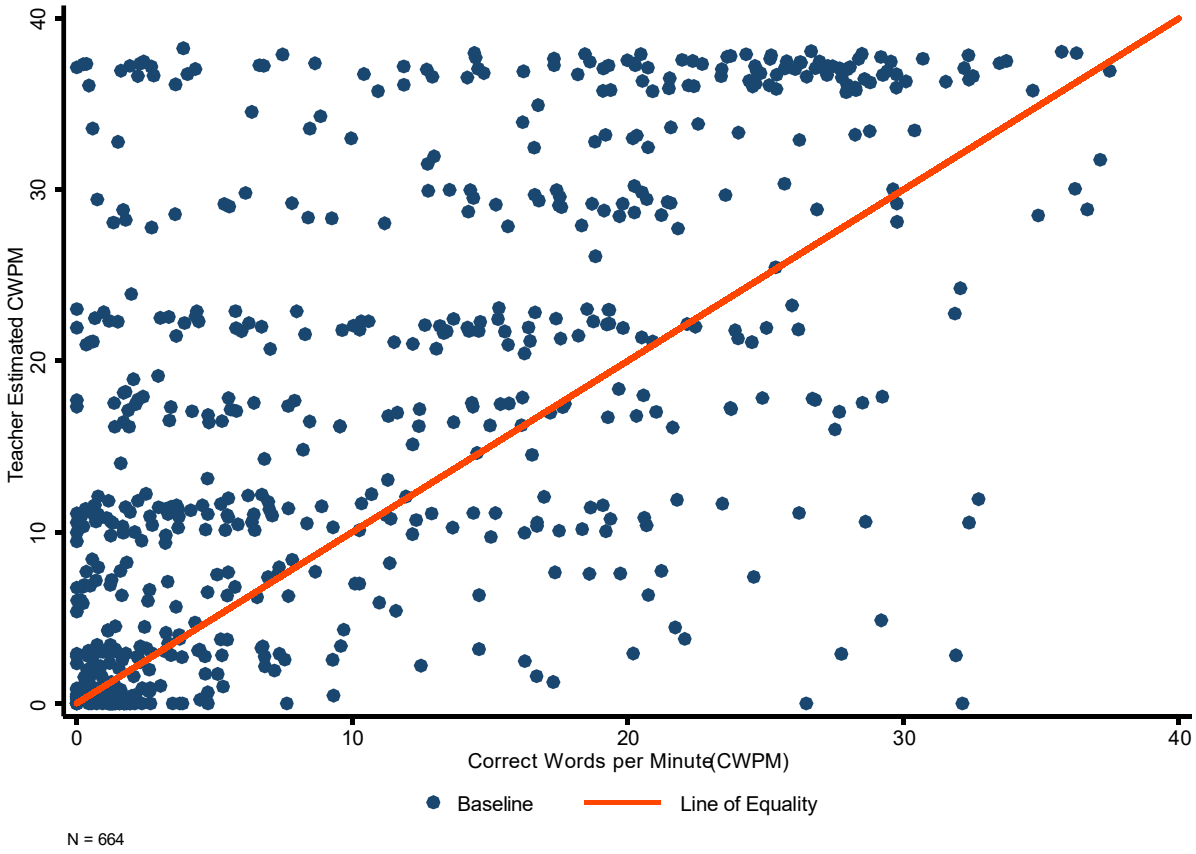
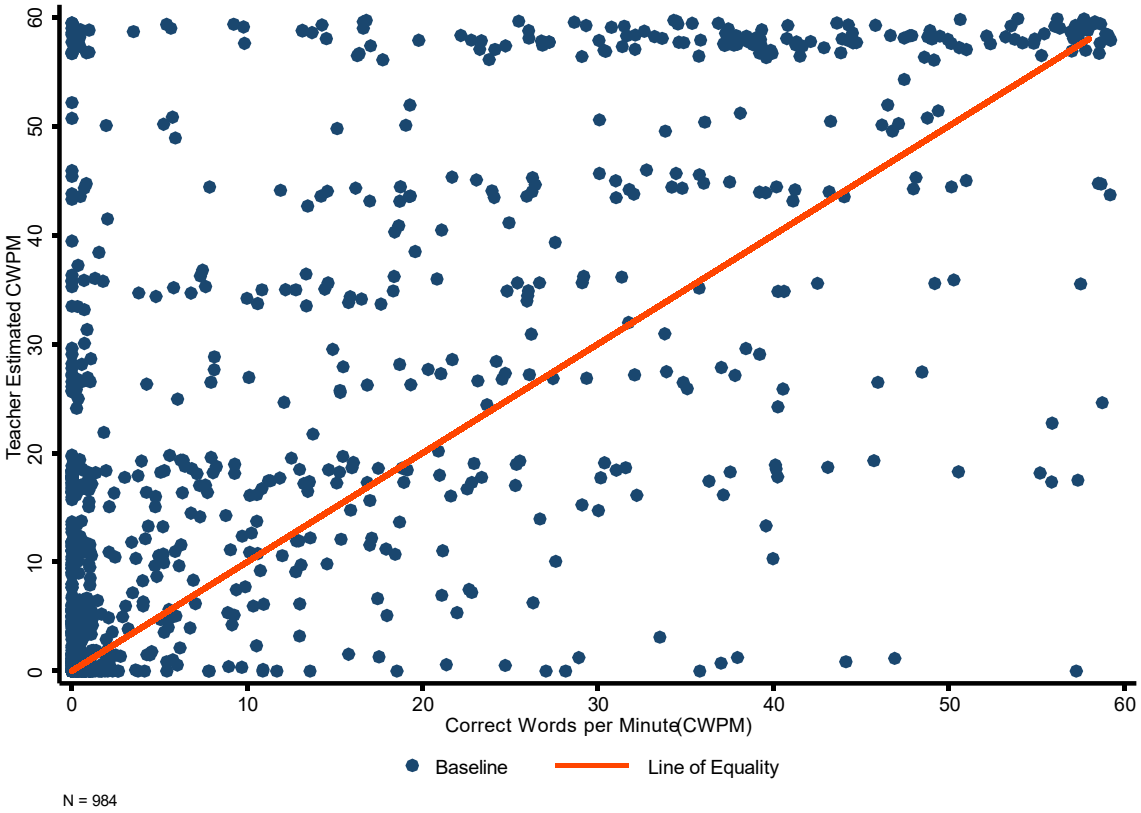
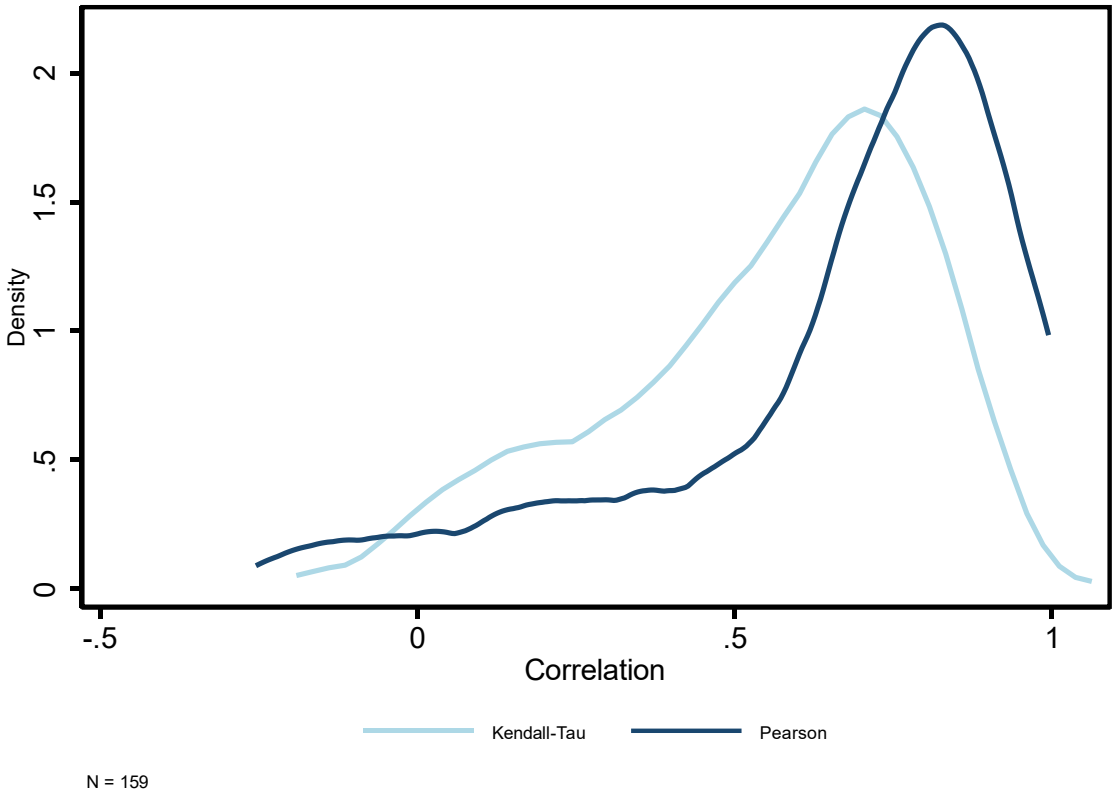


Figure 13: Teacher Estimate vs Learner Performance at Baseline – Setswana-Sesotho language



Summary measures of the alignment of teacher estimates and measured scores paint a somewhat clear picture. Figure 14 plots the distribution of both the Kendall-Tau and Pearson correlation measures at baseline. These measures differ based on the underlying aspect of judgement accuracy they aim to measure. The Kendall-Tau is the rank correlation and measures the degree to which teachers' estimation of relative reading proficiency agrees with the observed relative reading proficiency. Whereas the Pearson correlation measures the degree to which teacher estimates of absolute reading proficiency agree with the observed absolute reading proficiency of learners. The Kendall-Tau distribution is notably to the left of the Pearson distribution and the thickness of the tails show a relatively higher variance. The majority of teachers have a moderate positive correlation.

Figure 14: Correlation Measure Distributions



We next consider whether teachers are better able to identify learners at the extremes. Teachers were asked to select the strongest and weakest readers from the 10 learners that were randomly sampled for assessment in their class. We compare this against our ranking based on the EGRA. Given the non-trivial proportion of learners who scored zero on the oral reading fluency task (25% at baseline), ranks based on ORF scores would result in many ties. To overcome this, we first rank learners on their ORF task score and then break any ties in ranks by ranking them on their letter sound task score. At baseline, 46% of learners had at least one tie and at endline, 30% of learners were tied with at least one other learner. These learners were then ranked using their letter sound task scores and ties were broken using the letter sound task rank. Because the sample excludes learners who scored zero on the letter sound task, the probability of learners scoring the same on their letter sound tasks is low. For example, if two learners scored zero on the ORF task and received a tied rank of nine accordingly, the learner who scored higher on the letter sound task would retain their rank of nine and the other learner would be assigned a rank of 10. This measure of teacher knowledge of learner reading proficiency is somewhat crude as the question was not prescriptive that teachers should rank learners according to their oral reading fluency and they may have taken a broader view of how strong or weak a reader each learner is.

Figure 15 plots the actual rank according to the learner assessment for learners whom the teachers identified as the strongest readers at baseline. A teacher is considered to have correctly identified the strongest reader if the actual rank of the learner is one. Just under 50% of teachers correctly identified the learner who scored the highest in the oral reading fluency task as the strongest reader.

Figure 15. Rank of learners identified by teachers as strongest

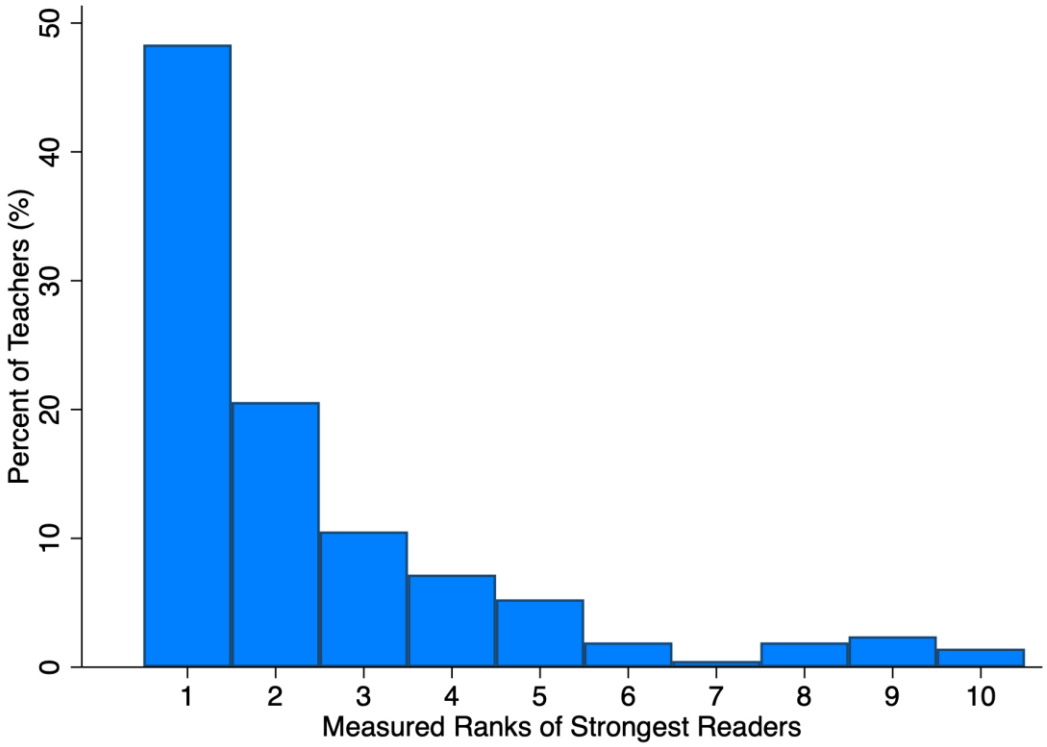
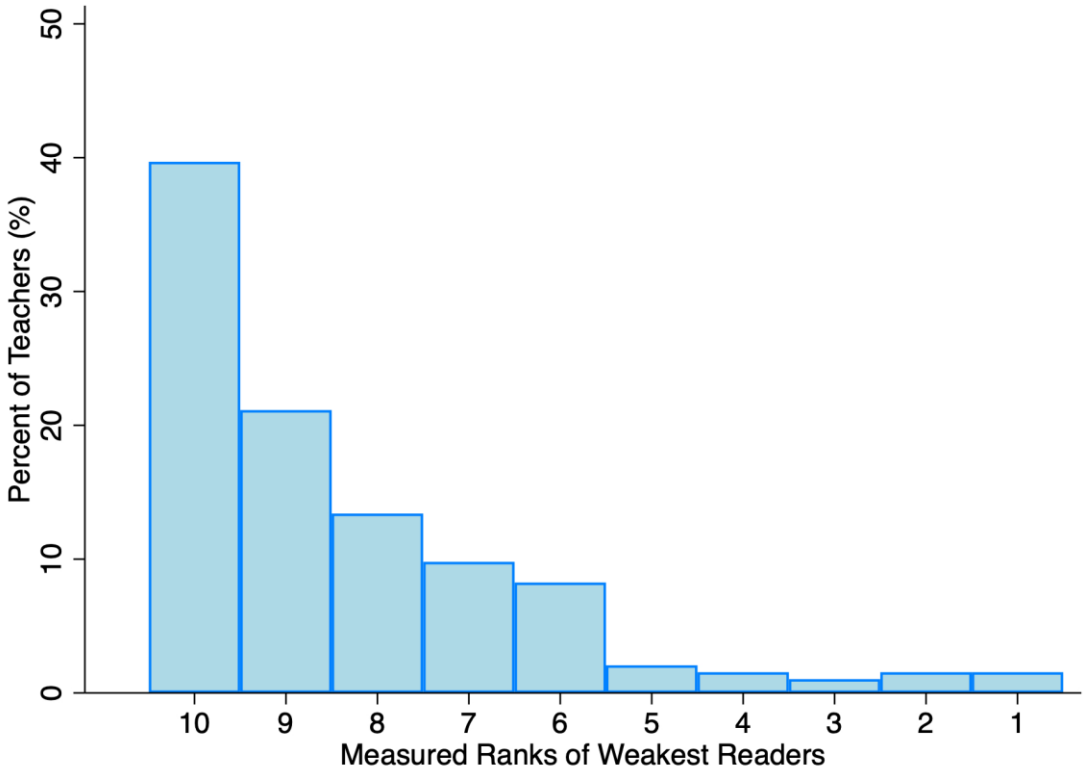


Figure 16 shows the ranks of the learners the teachers identified as the weakest readers. A smaller proportion of teachers correctly identified the lowest scoring learner as the weakest reader relative to the strongest reader.

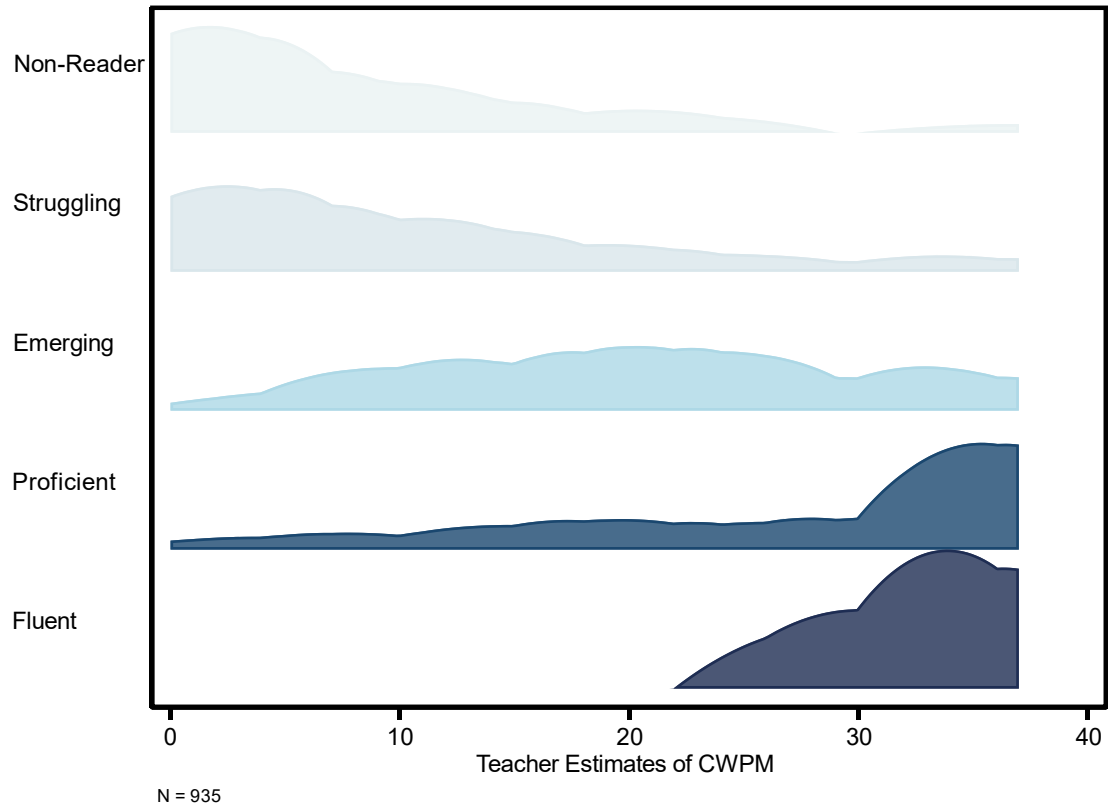
Figure 16. Rank of learners identified by teachers as weakest - Baseline



The relative rankings of learners are less important than knowledge of learners' substantive learning levels for practices such as ability grouping, differentiated instruction and remediation. We turn to the early-grade reading benchmarks to examine teachers' ability to classify learners according to these milestones. The benchmarks were established using large-scale EGRA data and aligned with meaningful points in learners' reading trajectories. Specifically, the grade 2 benchmark is aligned to the reading speed associated with a level of accuracy where learners can begin to shift cognitive effort from decoding to making meaning from text (Ardington et al., 2021). The grade 3 benchmark aligns with fluency levels that support comprehension and where the limiting factor becomes comprehension skills rather than reading fluency (Ardington et al., 2021). The benchmarks therefore correspond to qualitatively different levels of reading proficiency that require different instructional foci. For effective pedagogy, teachers need to know where their learners are located with respect to these benchmarks.

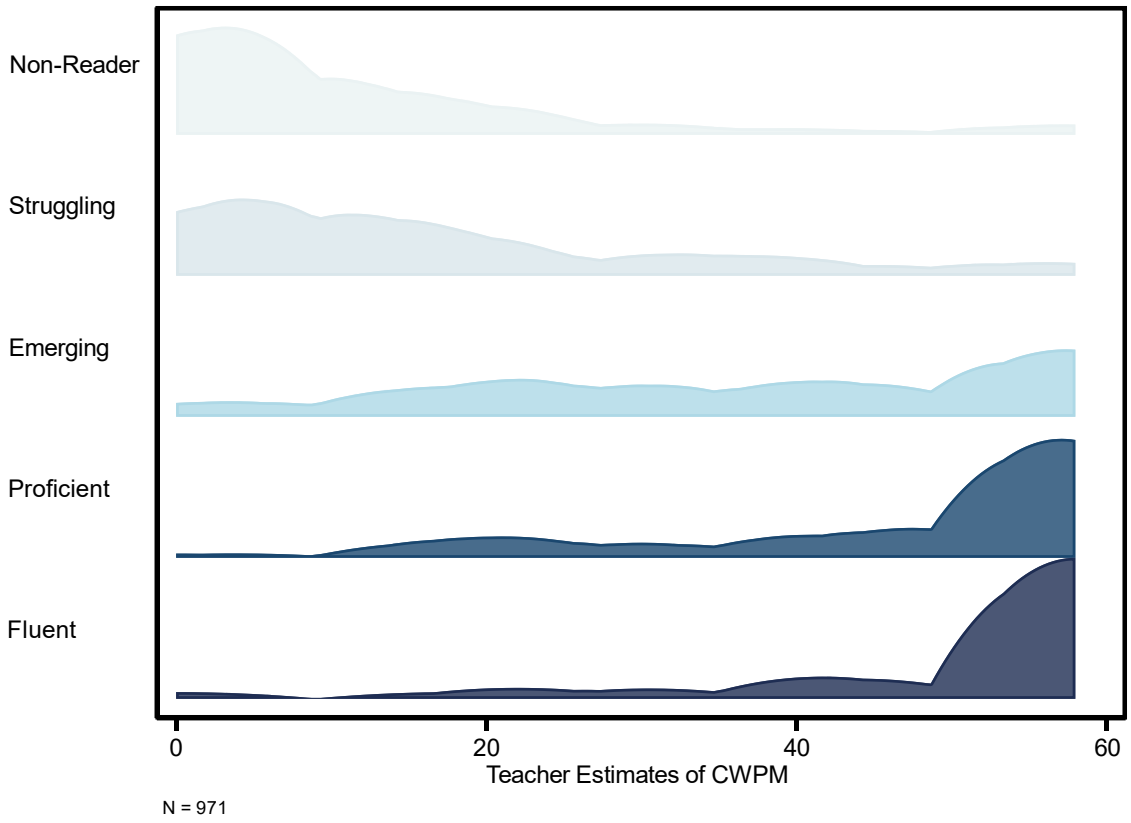
We classify learners into benchmark categories using the grade 2 thresholds for simplicity and compare this with the teacher estimates of the learners' benchmark category implied by their estimate of each learner's CWPM. Figure 17 and Figure 18 plot the distribution of teacher estimates for learners who fall within each benchmark level of reading proficiency.

Figure 17: Teacher Estimates of Learner Benchmark Category - Nguni



The distribution of teacher estimates for non-readers and struggling readers are nearly identical although the latter has a slightly higher variance. This suggests that teachers struggle to differentiate between Non-Readers and Struggling readers but they understand that these are the weaker learners. The distribution with the most variance in teacher estimates is the emerging reader category which suggests that this is the most difficult group for teachers to classify correctly. Finally, the distribution for proficient readers has a long tail to the left which is consistent with teachers' over-estimations of learners' reading abilities evident in Figure 12 and Figure 13. We observe similar patterns in the distributions of teacher estimates by benchmark category for Sesotho-setswana learners in Figure 18.

Figure 18: Teacher Estimates of Learner Benchmark Category - Sesotho-Setswana



For each teacher, we calculate the proportion of their 10 learners that they correctly classified according to the grade 2 reading benchmark thresholds. The distribution of these scores is shown in Figure 19. The average proportion of learners that teachers correctly classified is 37% which is four of the 10 sampled learners.

Figure 19. Distribution of proportion of learners correctly classified

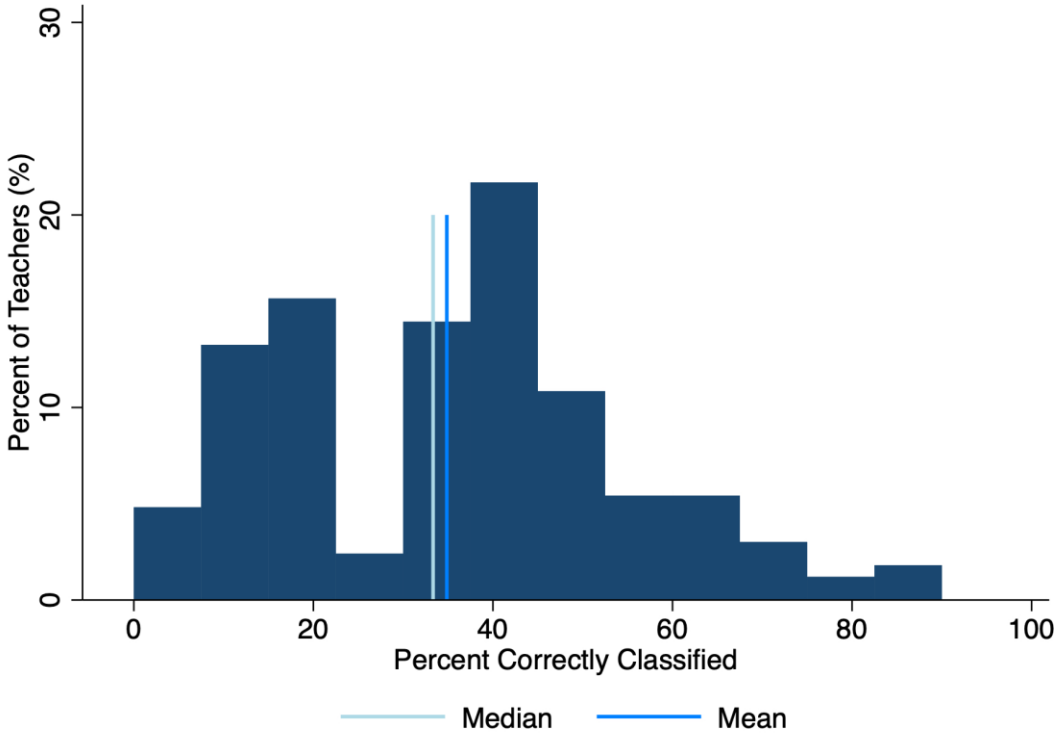
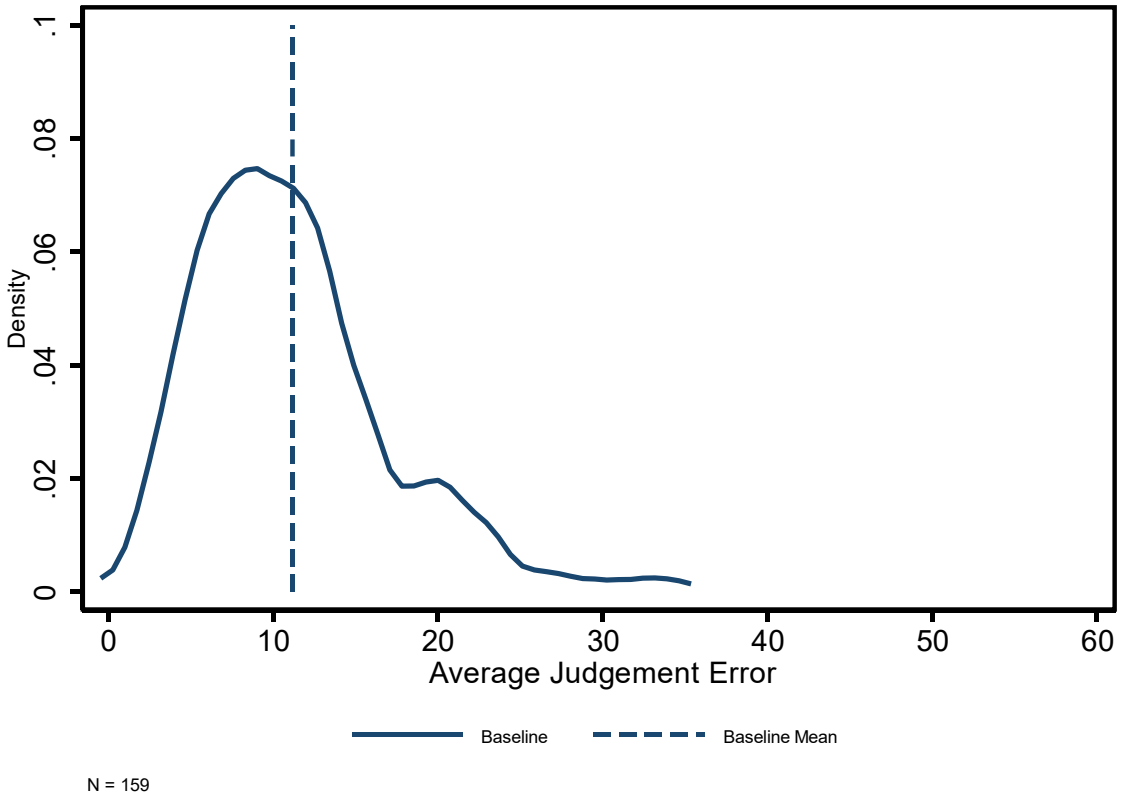


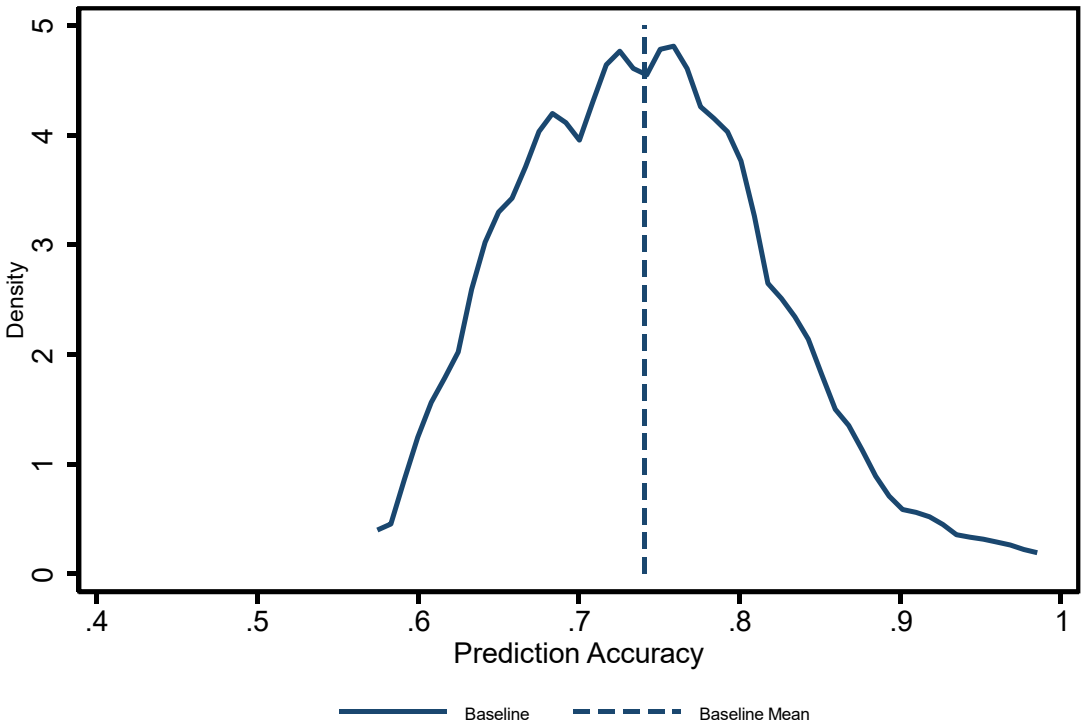
Figure 20 plots the distribution of average judgement error at baseline. The distribution is right-skewed and peaks around the mean of 10 words average judgement error. On average, teachers misestimate the CWPM of their learners by 12 words. In benchmark terms, this would correspond to a teacher estimating that a learner falls into a category one level higher than they do. Each category has distinct learning needs associated with it and hence different meanings for how a teacher would address those needs in their instructional approach. It is therefore crucial that teachers are able to classify learners correctly into benchmark categories if they are to meet the learning needs of learners.

Figure 20: Average Judgment Error Distribution



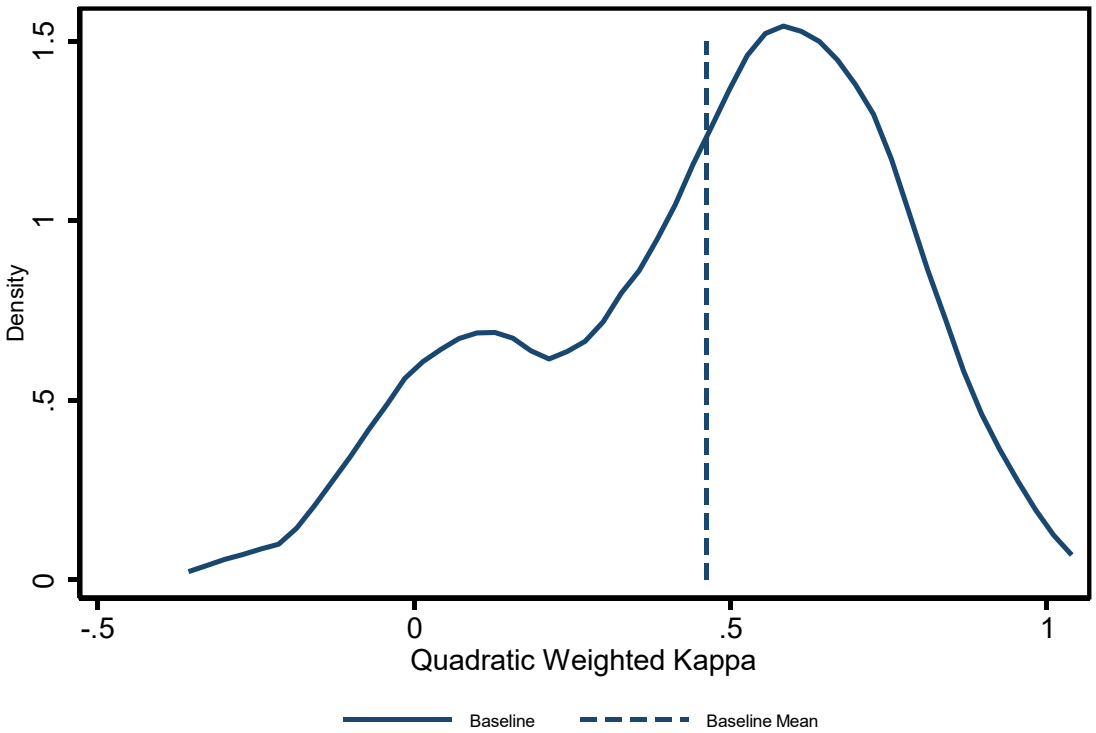
We further explore how well teachers are able to classify learners into benchmark categories using the prediction accuracy and quadratic weighted kappa measures, plotted in Figure 21 and Figure 22 respectively. The location of the prediction accuracy distribution, which is between 0.5 and 1, suggests that teachers' prediction accuracy in terms of distinguishing between the benchmark categories of learners, is high. At baseline, the average prediction accuracy was roughly 0.74 and the variance was lower than what we would expect with a normally distributed outcome. The distribution of the quadratic kappa measure which accounts for the size of differences in predictions by weighting these differences, shows larger variance. The average kappa score of 0.46 corresponds to a fair concordance of teachers' estimates.

Figure 21: Prediction Accuracy Distribution



N = 159

Figure 22: Quadratic Weighted Kappa Distribution



N = 159

Overall, our findings are consistent with the literature which suggests that in the context of large heterogeneous classes where communalized pedagogies are dominant, individualized assessment is limited and teacher knowledge of their learners' reading proficiency is sub-optimal. We find evidence that teachers overestimate the performance of their learners regardless of where these learners fall on the achievement distribution. Further, the size of misestimation is large and on average corresponds to teachers misclassifying their learners by at least one benchmark category level. We also find large variation of the teacher estimates even within a single category level. The extent to which the summary outcome measures are able to pick up some of the underlying variation in teacher estimates differs with some measures like the Kendall-tau rank correlation and quadratic weighted kappa performing better than others. In the next section, we first conduct statistical power analysis to investigate our ability to estimate the effects of the intervention using the outcome measures and our identification strategy before turning to examine the impact of the intervention.

7.2. Evaluation of Pilot Intervention

7.2.1. Quality of implementation

Program compliance is an issue commonly measured in impact evaluations of development programs. This is where participants are given access to the intervention but may choose for whatever reason to not implement or apply it. In our study, even in cases where teachers may use the intervention, they may do so inappropriately or inadequately. Brodie et al. (2002) investigate in-service teacher take-up of a learner-centred teaching practice program and find that primary school teachers from under-resourced schools took up the program in form but not in substance. In the context of this intervention, the form is to conduct assessments at set intervals, but the substance is that they engage with the process and outcomes of those assessments such that they update their knowledge of learning levels and adapt their instruction accordingly. Poor take-up potentially dilutes the intended treatment effect and limits the extent to which it can be identified, where it does exist.

Table 17 describes intervention take-up rates for this study. Of the 127 teachers who received the intervention, 86 teachers confirmed that they attended training during the endline interview. Twenty-nine teachers said they had not attended the training despite the training register data saying otherwise and a further 10 said they had attended some other assessment training unrelated to this intervention. It is possible that the teachers did not interpret the question to include our training, but fieldworkers were instructed to specifically prompt about our training.

Amongst those who confirmed attending training, 75 said that they had used the intervention materials and six teachers declined to answer the question. Four of the seven teachers who said they did not use the intervention materials said they did not receive them or that the materials were incomplete despite materials being distributed to all teachers.

Table 17: Take-Up Rates of Intervention among trained teachers

		Frequency	Proportion
Training attendance			
	Confirmed attendance	86	44%
	Attended other training	37	19%
	Attended no training	74	38%
	Total	**197**	
Use of materials for those confirming attendance			
	Yes	75	87%
	No	7	8%
	Refused	4	5%
	Total	**86**	
Progress chart use rating for those using materials			
	No use	4	5%
	Low use	20	27%
	High use	51	68%
	Total	**75**	

Note: Sample excludes the school where no teachers attended training.

We further classify the teachers who confirm using the intervention materials into use categories based on the photographs taken of their progress charts at endline. Sixty-eight (68) percent of these teachers had used the progress charts at the desired frequency which was three to four rounds of assessments. In the analyses that follow, we distinguish between trained teachers for whom we have evidence of reasonable uptake (high use) and the other trained teachers. The sample sizes and proportions are shown in Table 18.

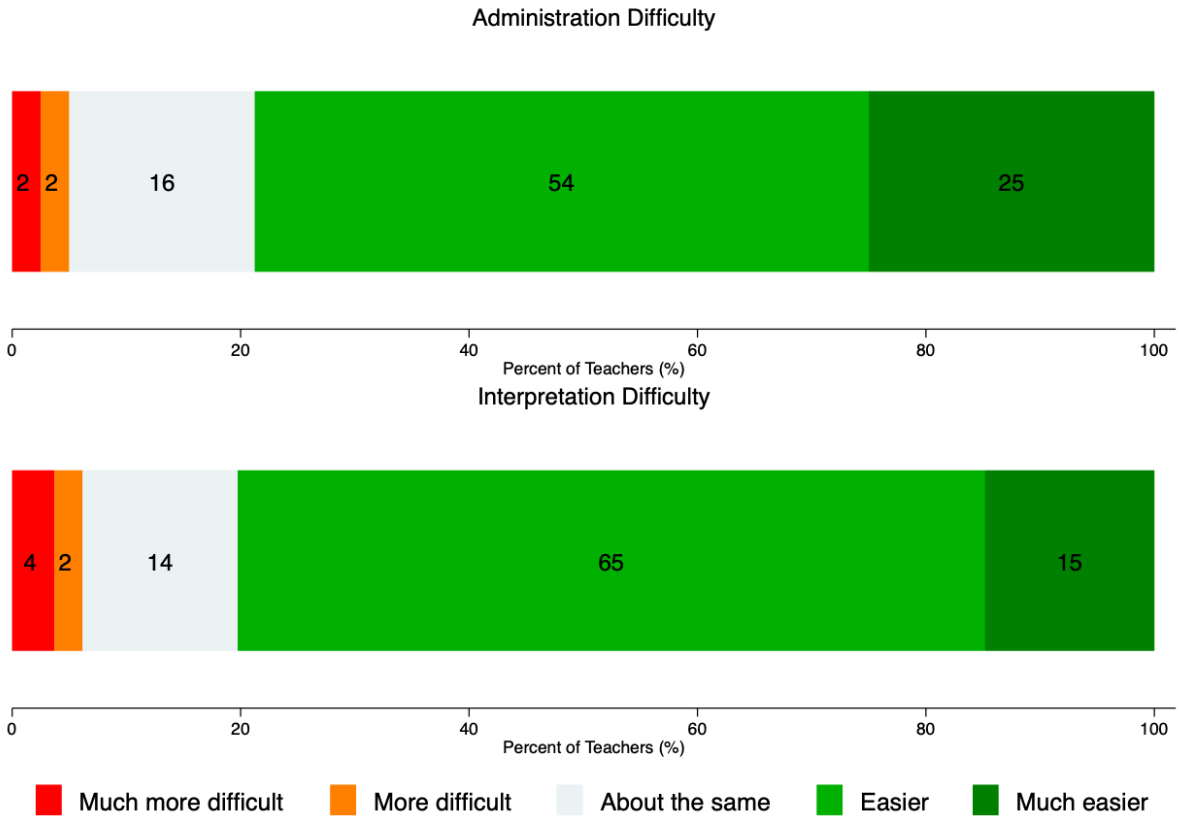
Table 18: Trained teacher categories

	Frequency	Proportion
Trained – high use	51	26%
Trained – other	80	41%
Control	66	33%
Total	**197**	

Note: Sample excludes the school where no teachers attended training.

Trained teachers were asked to rate the difficulty of administration and interpretation of the intervention assessment relative to the traditional EGRA (Figure 23). The vast majority of teachers report that the intervention assessment is both easier to administer and easier to interpret.

Figure 23: Teacher Rating of Intervention Administration and Interpretation Difficulty



7.2.2. Estimated Impact on Judgement Accuracy

As discussed in section 0, we employ three strategies to estimate the impact of the intervention on teacher judgement accuracy. These are the intent-to-treat (ITT) average treatment effect, the local average treatment effect (LATE) and the average treatment effect on the treated (ATT). The definitions of treatment under these three strategies differ. Under ITT, treatment refers to random assignment to receive the invitation to training. Under the ATT and LATE, treatment refers to whether a teacher was trained or not, but the latter strategy uses random assignment as an instrument to isolate the effect of the training. We estimate the ITT by fitting a multivariate OLS model, the LATE by fitting a multivariate two-stage least square (2SLS) model and the ATT by fitting a difference-in-difference (DID) model. The estimated coefficients of interest for each outcome measure are reported in Table 19.

Table 19: Estimated Treatment Coefficients

	(1)	(2)	(3)	
Variables	ITT	LATE	ATT	N
Rank Correlation	0.0343	0.0418	0.0995**	197
	(0.0382)	(0.0397)	(0.0434)	
Pearson Correlation	0.0359	0.0437	0.0520	197
	(0.0431)	(0.0452)	(0.0540)	
Judgement Error	-0.417	-0.516	-0.235	197
	(0.783)	(0.831)	(1.020)	
Prediction Accuracy	0.00450	0.00554	0.0144	197
	(0.0115)	(0.0121)	(0.0144)	
Quadratic Weighted Kappa	0.0142	0.0176	0.0560	197
	(0.0377)	(0.0400)	(0.0448)	
Controls	Yes	Yes	No	

Note: Controls are strata fixed effects, years of experience and the baseline outcome. *, **, *** indicate significance at the 10-percent, 5-percent and 1-percent significance level respectively. Robust standard errors in parentheses.

The ITT estimate is not statistically significant across all outcome measures. The magnitude and direction of the coefficients indicate that being invited to receive the intervention is associated with a small increase in rank correlation, point correlation and prediction accuracy as well as a small decrease in judgement error. Similarly, the LATE estimate is not statistically significant across all measures although the coefficients are larger in magnitude than the ITT estimate. These estimates also indicate the intervention slightly increased rank correlation, point correlation and prediction accuracy amongst the complier sub-group. The intervention also slightly decreased judgement error. The ATT estimate for the rank correlation measure is statistically significant at the 5 percent level. The correlation between treatment teacher

estimated rank order and measured rank order improved by 0.1 points, *ceteris paribus*. The coefficients on the other outcome measures are not statistically significant but lead to similar conclusions as the ITT and LATE estimates.

Because this is a pilot study on a light-touch intervention, we are interested in the impact on of judgement accuracy when teachers use the intervention as intended. The analysis of the learner progress charts indicates variable use of intervention materials with no evidence of take-up for some teachers. We, therefore, restrict our sample to examine the impact of the intervention of assessment efficacy for the sub-sample of treatment teachers who were classified as high-use. In doing this, we are able to generate suggestive evidence of the impact of the intervention when it is implemented well. The counterfactual sample is then trimmed to control teachers in the same schools which results in a sample size of 28 schools and a total of 100 teachers.

Table 20 describes the average characteristics of teachers in the restricted sample by training status. The teachers in the restricted sample on average have similar characteristics compared to those in the full sample (Table 11) except for a lower proportion with a post-matric qualification. Moreover, there is no statistically significant difference in the baseline characteristics of the trained teachers and their untrained counterparts. The similarity of the two groups at baseline supports the plausibility of the parallel trends assumption underlying our DiD estimates in this restricted sample.

Table 20: Average Characteristics of Restricted Sample

	All		Trained – High Use		Untrained – Matching Schools		P-Value
Variable	Mean	SD	Mean	SD	Mean	SD	
Experience (Years)	16	10,23	16	9,14	17	11,31	0,49
Experience in Foundation Phase (Years)	13	9,60	12	8,34	14	10,82	0,55
Has a Diploma	0,31	0,46	0,29	0,46	0,33	0,47	0,73
Has a Degree	0,33	0,47	0,37	0,49	0,29	0,46	0,36
Training in Foundation Phase Teaching	0,70	0,46	0,75	0,44	0,65	0,48	0,32
EGRA Trained	0,46	0,50	0,39	0,49	0,54	0,50	0,13
Class Size	44,85	6,86	44,96	6,95	44,73	6,84	0,87
Average Learner ORF Score	12,76	7,04	12,50	6,87	13,02	7,28	0,72
Average Learner Letter Sounds Score	33,23	11,72	33,78	11,86	32,65	11,65	0,63
Groups Learners	0,83	0,38	0,80	0,40	0,85	0,36	0,48
Number of Teachers	100		51		49		

Note: SD = Standard Deviation. P-Value of the significance in difference between the treatment and control groups. A value of 0.05 or less indicates that the difference in the mean values of the groups is statistically significant or different from zero at the 5% significance level.

We estimate the treatment effects within the restricted sample using our preferred model, the DiD. The full model results are reported in Table 21. Using the Kendall-tau rank correlation, we find no statistically significant difference in the outcomes of control teachers between baseline and endline. The coefficient on the difference-in-difference term is significant at the 5% level and suggests that the correlation between the estimates of treatment teachers and learner performance improved by 0.11 points relative to their control counterparts between baseline and endline. This is slightly larger than the improvement of the full treatment group. Assuming that the DID strips out the selection bias of which kinds of teachers are more likely to use the intervention appropriately, it implies that under greater fidelity, there would have been a greater impact on their knowledge of learning levels. The coefficient of interest is not statistically significant under the other outcome measure model specifications, but the magnitudes are larger than under the full sample estimation.

Table 21: Difference-in-Difference Coefficients - Restricted Sample

Variables	(1) Rank Correlation	(2) Pearson Correlation	(3) Judgement Error	(4) Prediction Accuracy	(5) Quadratic Kappa
Endline	-0.00138	0.0312	-0.504	0.0103	0.0567*
	(0.0286)	(0.0370)	(0.653)	(0.0114)	(0.0320)
Treat x Endline	0.110**	0.0944	-0.547	0.000106	0.0595
	(0.0534)	(0.0657)	(1.412)	(0.0188)	(0.0558)
Constant	0.524***	0.621***	11.49***	0.736***	0.454***
	(0.0152)	(0.0187)	(0.400)	(0.00526)	(0.0158)
Observations	176	176	178	178	177
R-squared	0.101	0.095	0.017	0.015	0.125
Number of Teachers	100	100	100	100	100

Note: Models include individual fixed effects. *, **, *** indicate significance at the 10-percent, 5-percent and 1-percent significance level respectively. Standard errors in parentheses

8. Discussion

In a sample where just under 50% of schools have received EGRA training from DBE, we find considerable variability in teachers' exposure to the EGRA tool. Overall, 46% of teachers reported not having any exposure to EGRA with Mpumalanga having the highest proportion at 62% despite 9 of the 10 evaluation schools having received EGRA training from DBE. We also find that among those who had been trained or had experience in conducting EGRA, only 74% of teachers had completed an assessment by May which was halfway through the second term. According to CAPS, Grade 2 and 3 teachers need to have conducted at least one assessment per school term. We would hence expect all teachers in the Eastern Cape and Mpumalanga to have conducted the EGRA as these were the provinces where all but one school were part of the DBE EGRA rollout. What we find instead is that only 76% and 38% of teachers in the Eastern Cape and Mpumalanga, respectively, had conducted the assessment by May.

Amongst teachers who had been exposed to EGRA and had conducted at least one assessment by May, we find that on average, teachers take 5-10 minutes to complete an EGRA with a single learner. This implies that assessing the average class would take roughly four contact hours of the seven hours per week allocated to home language teaching and learning activities. Teachers reported making this "lost" instructional time up by assigning individual tasks to learners during assessments. The DBE workbooks are a key resource in enabling individual work during learner assessments. Another assessment strategy that

teachers employ is to split assessments over several days. These self-reported practices reflect positively on the ability of teachers to implement individualised assessments through the EGRA even in contexts where classes can be as large as 70 learners in a single class.

We find that teachers' understanding of the EGRA tool to be poor, specifically on what the tool measures. Thirty-five (35%) percent of teachers who had completed an assessment by May expressed that increasing the time limit from one minute would make it easier for learners to complete the assessment passage. This is despite receiving training that the task is timed to measure speed in addition to accuracy and learners are not meant to finish the passage. The assessment training provided by DBE may likely be too light-touch for teachers to fully understand the tool. This is also reflected in the proportion of teachers reporting not having received EGRA training, those who had not used the tool by May and those expressing the need for further assessment training.

In relation to baseline teacher judgement accuracy, we find that teachers tend to overestimate the performance of their learners and the size of the misestimation is quite large with the average judgement error being 12 words read correctly per minute. This corresponds to teachers misclassifying learners by at least one benchmark category. Learner performance improved over the period which resulted in a smaller proportion of learners who cannot read a single word, but the majority of learners still fall within the struggling reader benchmark category. This is also the benchmark category with the largest variation in teacher estimates which suggests that teachers struggle to classify these learners. This is consistent with our finding that teachers tend to do better at identifying the strongest readers than they do with identifying the weakest readers, even though they misestimate learner performance across the achievement distribution.

In terms of the summary measures, the estimated Pearson correlation and prediction accuracy suggest that teacher judgment is moderate. On the other hand, Kendall-tau and quadratic kappa measures show that teacher judgement accuracy is low to moderate. Our estimates of teacher judgement accuracy are consistent with those from other studies which found an average correlation of between 0.36 and 0.65 which corresponds to a range between weak and moderative positive correlation. Like van der Berg and Shepherd (2015), we argue that although our estimates fall within the moderate range of the general interpretation of these measures, the contextual interpretation is that teacher judgement accuracy is low. Teachers interact with learners regularly and are confronted by their learning levels daily. Finding 'moderate' teacher judgement accuracy highlights the need to improve the formative assessment process and teachers' interpretation of the results thereof because we would

expect teachers to have a strong sense of their learners' reading abilities given how much time they spend with learners.

Although we find that take-up of the intervention was low, our estimation of the impact of the intervention using the ITT, LATE and ATT parameters shows that the intervention was able to improve some aspects of teacher judgement accuracy. Namely, teacher knowledge of relative reading proficiency and the association between teacher estimates and learner performance. For the ITT and LATE, while the point estimates are positive (or negative in terms of judgement error), the coefficients are not statistically significant at conventional levels. We are unable to detect any impacts on teacher estimates of absolute reading proficiency or how well teachers are able to classify the sampled learners into benchmarking categories. We attribute this to the low take-up and use of the intervention. However, when we use the DID and focus on teachers attending training, we find evidence of shifts in teacher judgement accuracy. This is even more so when we restrict the sample to teachers with evidence of use and their untrained counterparts within the same school. One limitation of the analysis conducted in this study is the lack of statistical power. Although this was a small-scale pilot study, any efforts to take this work forward would need to carefully consider how to encourage fidelity to the intervention in a sample large enough to detect intervention impact where it does exist.

9. Conclusion

This study contributes initial evidence on how the newly established benchmarks can be used productively in classrooms and to the best of our knowledge some of the first quantitative evidence on teacher assessment efficacy in the Foundation Phase in South Africa. Together with the Department of Basic Education, we designed a simplified version of the Early Grade Reading Assessment and piloted this along with training on benchmarks and assessment resources to schools in the Eastern Cape, North West, Limpopo and Mpumalanga over the 2022 school year.

Contrary to what is captured in the DBE database, we find that just under half of teachers had no exposure to the EGRA which we argue was a large contributing factor to variable fidelity to the intervention. The intervention was designed as an enhancement to existing EGRA assessment practices and not to introduce EGRA to teachers. The training provided would have not been enough for teachers to feel comfortable with the assessment without the initial training from the DBE. We also find that amongst teachers who had received the DBE training or were experienced in conducting EGRA, their understanding of the tool was poor and some

teachers highlighted a need for further training and support. The training provided to teachers by DBE as part of the EGRA rollout may be too light-touch for teachers to fully understand the tool and how to implement it in their classrooms. This is one potential explanation for why 26% of teachers with exposure to EGRA had not conducted a single EGRA assessment by halfway through the second school term.

Given the findings on assessment practices, it is unsurprising that we find teacher judgement accuracy to be low. Particularly given the evidence suggesting that teachers have a poor understanding of formative assessment in general. Teachers tend to overestimate the performance of their learners across the achievement distribution and the size of the misestimation is large. On average, teachers misestimated the number of words that learners could read correctly by 12 words at baseline and correctly classified just under four of 10 learners into the benchmark categories. Teachers also had a moderate positive average rank correlation. If teachers are not aware of the learning levels of their learners, they cannot tailor their instruction in a way that meets their learning needs. Low teacher judgement accuracy is symptomatic of inefficiencies in the assessment process. Our findings suggest that targeted or differentiated instruction programs are unlikely to realize their full potential in the South African context if special attention is not paid to assessment practices. Teachers would likely be pitching their instruction above the level of learners even if they were to differentiate. Therefore, formative assessment practices need to be strengthened, and teacher knowledge of learning levels improved through teacher support. Such support will need to take into account the very difficult contexts in which teachers assess with overcrowded classrooms and a lack of basic assessment resources like text. In very large classes, one-on-one orally administered assessments can crowd out instructional time.

The intervention was designed recognising these difficult contexts by simplifying the EGRA, minimising the time taken to assess a single learner and providing teachers with the necessary resources. We find evidence that the intervention improved teachers' knowledge of relative reading proficiency, but we are unable to conclude on its impact on other aspects of teacher judgement accuracy. This does not mean that there was no impact, but rather that this small pilot study was statistically underpowered to detect effects even when they exist. There was also low take-up of the intervention and we only found effects when we focused on teachers who attended the training. The feedback received from teachers does indicate, however, that the overwhelming majority of them found it to be an improvement over the traditional EGRA in that it was easier to interpret and administer.

We, therefore, conclude that using reading benchmarks to enhance the assessment process is a low-cost and scalable way to disseminate benchmarks for productive classroom use. Given the weak understanding of formative assessment in general, evidence of the low impact of DBE training and limited take-up of our intervention, more intense training and support is required for teachers. Our findings underscore the importance of enhancing the resources, training, and support provided to teachers for the formative assessment process. Furthermore, the inclusion of the reading benchmarks could be useful in augmenting such support. Further research should focus on evaluating the intervention in a larger sample to increase statistical power to consider the effect, if any, on learner performance and heterogenous treatment effects.

References

Angrist, J.D. & Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Ardington, C., Wills, G., Pretorius, E., Deghaye, N., Mohohlwane, N., Menendez, A., Mtsatse, N. and van der Berg, S., 2020. Benchmarking early grade reading skills in Nguni languages.

Ardington, C., Wills, G., Pretorius, E., Mohohlwane, N. and Menendez, A., 2021. Benchmarking oral reading fluency in the early grades in Nguni languages. *International Journal of Educational Development*. 84. p.102433.

Banerjee, A.V., Cole, S., Duflo, E. & Linden, L. 2007. Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*. 122(3):1235–1264. DOI: 10.1162/qjec.122.3.1235.

Bassi, M., Meghir, C. & Reynoso, A. 2020. Education quality and teaching practices. *The Economic Journal*. 130(631):1937–1965. DOI: 10.1093/ej/ueaa022.

Brenner, H. & Kliebsch, U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*. 7(2):199–202. DOI: 10.1097/00001648-199603000-00016.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. 2015. The diagnostic skills of mathematics teachers. In *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand, Eds. Berlin: Springer.

Cilliers, J., Fleisch, B., Prinsloo, C. & Taylor, S. 2018. How to improve teaching practice? experimental comparison of centralized training and in-classroom coaching. *Stellenbosch Economic Working Papers*. 2018/WP15. DOI: 10.35489/bsg-rise-wp_2018/024.

Cunningham, S. 2021. *Causal Inference: The Mixtape*. New Haven: Yale University Press.

Department of Basic Education. 2011a. *National Curriculum Statement (NCS): Curriculum and Policy Statement (CAPS) Grades R-3: Home Languages*. Department of Basic Education: Pretoria.

Department of Basic Education. 2011b. *National Protocol for Assessments Grades R – 12*. Department of Basic Education: Pretoria.

References

Department of Basic Education. 2023. *PIRLS 2021: South African Preliminary Highlights Report.* Department of Basic Education: Pretoria.

Djaker, S., Ganimian, A. & Sabarwal, S. 2022. Primary- and middle-school teachers in South Asia overestimate the performance of their students. [Manuscript]

Fleisch, B. & Dixon, K. 2019. Identifying mechanisms of change in the early grade reading study in South Africa. *South African Journal of Education*. 39(.):1–12. DOI: 10.15700/saje.v39n3a1696.

Fleisch, B. 2018. *The Education Triple Cocktail: System-wide instructional reform in South Africa*. Cape Town: UCT Press.

Gamaroff, R. 2000. Rater reliability in language assessment: The bug of all bears. *System*. 28(1):31–53. DOI: 10.1016/s0346-251x(99)00059-7.

Govender, R. & Hugo, A.J. 2020. An analysis of the results of literacy assessments conducted in South African Primary Schools. *South African Journal of Childhood Education*. 10(1). DOI: 10.4102/sajce.v10i1.745.

Grandini, M., Bagli, E. and Visani, G., 2020. Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Hadjidemetriou, C. & Williams, J. 2001. Evaluating teachers' knowledge in relation to their children's learning. In *Proceedings of the British Society for Research into Learning Mathematics*. 1st ed. V. 21.

Heckman, J. 1997. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*. 32(3):441. DOI: 10.2307/146178.

Hoge, R.D. & Coladarci, T. 1989. Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*. 59(3):297. DOI: 10.2307/1170184.

Imbens, G.W. & Angrist, J.D. 1994. Identification and estimation of local average treatment effects. *Econometrica*. 62(2):467. DOI: 10.2307/2951620.

Imbens, G.W. & Rubin, D.B. 2015. *Causal inference in statistics, social, and Biomedical Sciences: An introduction*. Cambridge: Cambridge University Press.

References

Kahn-Lang, A. & Lang, K. 2019. The promise and pitfalls of differences-in-differences: Reflections on “16 and pregnant” and other applications. *Journal of Business & Economic Statistics*. 38(3):613–620. DOI: 10.1080/07350015.2018.1546591.

Kaiser, J., Südkamp, A. & Möller, J. 2017. The effects of student characteristics on teachers’ judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology*. 109(6):871–888. DOI: 10.1037/edu0000156.

Kanjee, A. & Bhana, J. 2022. Lessons learner and evidence of impact: Formative assessment in an integrated reading and mathematics intervention. In *Early grade reading and mathematics interventions in South Africa*. N. Spaull & S. Taylor, Eds. Cape Town, South Africa: Oxford University Press Southern Africa (Pty) Limited.

Kanjee, A. & Croft, C. 2012. Annual Meeting of the American Educational Research Association. In *Examining Assessment for Learning in the Schools*. Vancouver, British Columbia, Canada. Available: https://www.aera.net/Publications/Online-Paper-Repository/AERA-Online-Paper-Repository.

Kanjee, A. & Mthembu, J. 2015. Assessment Literacy of Foundation Phase Teachers: An exploratory study. *South African Journal of Childhood Education*. 5(1):26. DOI: 10.4102/sajce.v5i1.354.

Kanjee, A. & Sayed, Y. 2013. Assessment policy in post-apartheid South Africa: Challenges for improving education quality and learning. *Assessment in Education: Principles, Policy & Practice*. 20(4):442–469. DOI: 10.1080/0969594x.2013.838541.

Kanjee, A. 2020. Exploring primary school teachers’ use of formative assessment across fee and no-Fee Schools. *South African Journal of Childhood Education*. 10(1). DOI: 10.4102/sajce.v10i1.824.

Kaufmann, E. 2020. How accurately do teachers’ judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology*. 63:101902. DOI: 10.1016/j.cedpsych.2020.101902.

Kilday, C.R., Kinzie, M.B., Mashburn, A.J. & Whittaker, J.V. 2012. Accuracy of teacher judgments of preschoolers’ math skills. *Journal of Psychoeducational Assessment*. 30(2):148–159. DOI: 10.1177/0734282911412722.

References

Kolovou, D., Naumann, A., Hochweber, J. & Praetorius, A.-K. 2021. Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education*. 100:103298. DOI: 10.1016/j.tate.2021.103298.

Maunganidze, L., Ruhode, N., Shoniwa, L., Nyanhongo, S., M. Kasayira, J. & Sodi, T. 2008. Teacher ratings and standardised test scores: How good for predicting achievement in students with learning support placement? *Journal of Psychology in Africa*. 18(2):255–258. DOI: 10.1080/14330237.2008.10820194.

Menendez, A. & Ardington, C. 2018. (Rep.). *Impact Evaluation of USAID/South Africa Story Powered School Program – Baseline.* Technical Report.

Mkhwanazi, H.N., Joubert, I., Phatudi, N.C. & Fraser, W.J. 2014. Teachers' use of formative assessment for the teaching of reading comprehension in grade 3. *Mediterranean Journal of Social Sciences*. DOI: 10.5901/mjss.2014.v5n7p468.

Mohohlwane, N., Wills, G. & Ardington, C. 2022. A review of recent efforts to benchmark early reading skills in South African languages. In *Early Grade Reading in South Africa*. N. Spaull & E. Pretorius, Eds. Oxford University Press.

Moon, T.R. 2005. The role of assessment in differentiation. *Theory Into Practice*. 44(3):226–233. DOI: 10.1207/s15430421tip4403_7.

Motilal, G.B. & Fleisch, B. 2020. The triple cocktail programme to improve the teaching of reading: Types of engagement. *South African Journal of Childhood Education*. 10(1). DOI: 10.4102/sajce.v10i1.709.

Piper, B., Simmons Zuilkowski, S., Dubeck, M., Jepkemei, E. & King, S.J. 2018. Identifying the essential ingredients to literacy and numeracy improvement: Teacher Professional Development and coaching, student textbooks, and structured teachers' guides. *World Development*. 106:324–336. DOI: 10.1016/j.worlddev.2018.01.018.

Ramaphosa, C. 2019. State of the Nation Address South Africa.

Ready, D.D. & Chu, E.M. 2015. Sociodemographic inequality in early literacy development: The role of teacher Perceptual Accuracy. *Early Education and Development*. 26(7):970–987. DOI: 10.1080/10409289.2015.1004516.

Spaull, N. 2013. Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*. 33(5):436–447. DOI: 10.1016/j.ijedudev.2012.09.009.

References

Spaull. N. 2023. *2023 Background Report for the 2030 Reading Panel*. Cape Town. Access URL: https://www.readingpanel.co.za/_files/ugd/b385b7_7476724ee8a74ba8be8320a3be46b5cc.pdf

Spaull, N., Courtney, P. & Qvist, J. 2022. Mathematical Stunting in South Africa: An analysis of Grade 5 mathematics outcomes in TIMSS 2015 and 2019. In *Early Grade Mathematics in South Africa*. N. Roberts & H. Venkat, Eds. Cape Town: Oxford University Press.

Spaull, N. & Pretorius, E. 2022. Coming or going? The prioritisation of early grade reading in South Africa. In *Early grade reading in South Africa*. N. Spaull & E. Pretorius, Eds. Cape Town, South Africa: Oxford University Press Southern Africa (Pty) Limited.

Südkamp, A., Kaiser, J. & Möller, J. 2012. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*. 104(3):743–762. DOI: 10.1037/a0027627.

Taylor, S., Cilliers, J., Prinsloo, C., Fleisch, B. & Reddy, V. 2017. The Early Grade Reading Study: Impact evaluation after two years of interventions. *Improving Early Grade Reading in South Africa*. DOI: https://www.jet.org.za/clearinghouse/projects/primted/resources/language-and-literacy-resources-repository/egrs-technical-report-13-oct-2017.pdf.

Van der Berg, S. & Shepherd, D. 2015. Continuous Assessment and Matriculation Examination Marks – an empirical examination. *South African Journal of Childhood Education*. 5(2):17. DOI: 10.4102/sajce.v5i2.391.

Van der Berg, S. 2015. What the annual national assessments can tell us about learning deficits over the education system and the school career. *South African Journal of Childhood Education*. 5(2). DOI: 10.4102/sajce.v5i2.381.

Van der Berg, S., Taylor, S., Gustafsson, M., Spaull, N. & Armstrong, P. 2011. Reort for the National Planning Commission: Improving education quality in South Africa. University of Stellenbosch

Van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., van Merriënboer, J. & Visscher, A.J. 2018. Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement*. 30(1):51–67. DOI: 10.1080/09243453.2018.1539013.

Wiliam, D. & Thompson, M. 2017. Integrating assessment with learning: What will it take to make it work? *The Future of Assessment*. 53–82. DOI: 10.4324/9781315086545-3.

References

Wills, G., Ardington, C. & Sebaeng, M.L. 2022. Foundational skills in home language reading in South Africa: Empirical evidence from 2015–2021. In *Early Grade Reading in South Africa*. N. Spaull & E. Pretorius, Eds. Oxford University Press.

Yilmaz, A.E. and Demirhan, H. 2023. Weighted kappa measures for ordinal multi-class classification performance. Applied Soft Computing, 134, p.110020.