EGRS
EARLY GRADE READING STUDY

The Second Early Grade Reading Study

Year 3 Report

Evidence after three years of implementation

education
Department of Education
REPUBLIC OF SOUTH AFRICA

education
MPUMALANGA PROVINCE
REPUBLIC OF SOUTH AFRICA

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

USAID
FROM THE AMERICAN PEOPLE

Contents

List of Figures

List of Tables

Acronyms

CAPS	Curriculum and Assessment Policy Statement
EFAL	English as First Additional Language
EGRS	Early Grade Reading Study
EGRS II	Second Early Grade Reading Study
HL	Home Language
HOD	Head of Department
ICT	Information Communication Technology
LoLT	Language of Learning and Teaching
LTSM	Learning and Teaching Support Material
NECT	National Education Collaboration Trust
RAN	Rapid Automised Naming
RCT	Randomised Control Trial
SMT	School Management Team
WCPM	Words Correct Per Minute

Executive Summary

Over the past eight years, the Early Grade Reading Studies have been advancing knowledge on the system-wide improvement of early grade reading in both African languages and English as a second language in South Africa (Fleisch, 2018 & 2019, Cilliers, 2019a&b; Kotze, 2019). Much of this new knowledge centres on the effectiveness and the mechanisms of a basic structured pedagogic change model - a combination of detailed daily lesson plans, high-quality educational materials and centralised training/on-site coaching. This research shows that the approach is an effective and sustainable way of improving early grade reading teaching system-wide.

While there is general agreement that instructional coaching is a key component of a structured pedagogy program approach, concerns have been raised both about the relatively high cost of on-site instructional coaching and the size of the pool of high-quality coaches available particularly in rural areas. To address these concerns, the Second Early Grade Reading Study (EGRS II) investigated the viability and cost-effectiveness of an alternative model of instructional coaching, what we have come to refer to as ‘virtual’ coaching. Virtual coaching makes use of a mixture of tablet technology, short video-clips, text messaging and cellular voice conversations. Preliminary results (Kotze, 2019) from the first year of the intervention suggested that both on-site and virtual coaching were equally effective in improving early grade reading outcomes in English as a second language, but that these gains were for basic literacy skills, particularly oral word recognition in the second language.

If these results hold throughout the Foundation Phase, it will mean that teachers can be supported remotely, which would lower transport costs, increase the number of schools that can be served by one coach, and reduce the reliance on finding enough quality coaches residing in targeted areas. However, there are reasons why virtual coaching and electronic lesson plans could be less effective. Teachers might struggle to adapt to using new technology, and require substantial training up-front to use the technology. Moreover, face-to-face engagement might be necessary to build a relationship of trust between the teacher and the coach, which allows the teacher to be vulnerable and discuss ways to improve her teaching. A lack of face-to-face engagement could also lead to less accountability, all of which means that such an intervention relies on a greater degree of self-motivation from teachers.

We address this question by experimentally comparing on-site with virtual coaching in the context of teaching English as a second language in South Africa. In both programmes, teachers received the same learning aids and training at the start of the programme, and the curriculum and content of lesson plans were the same. In 50 randomly assigned schools teachers received on-site in-classroom visits by coaches about 12 times a year, supplemented with needs-based clustered workshops, and the daily lesson plans were paper-based. In another 50 other schools, teachers received virtual coaching through phone calls, text messages and WhatsApp groups; and the lesson plans were on an electronic tablet. An additional 80 schools were assigned to the control.

The programmes were implemented over three years, with successive grade implementation (grade one teachers in the first year, grade two teachers in the second year, etc). We tracked the same cohort of learners over three years, starting in February 2017 when they entered grade one, and ending in November 2019. This report provides results after the full three years of program implementation. Two main outcomes of interest were targeted: oral language proficiency in English, and reading proficiency in English.

There are three main results to highlight. The main finding is that the on-site coaching program was more successful at improving the intended learning outcomes than the virtual coaching intervention. On-site coaching had statistically significant positive impacts on both English oral language proficiency (0.36 standard deviations) and English reading proficiency (0.18 standard deviations). In contrast, the virtual coaching program only improved English oral language proficiency by 0.15 standard deviations — less than half the magnitude, relative to on-site coaching — and had no statistically detectable impact on reading proficiency skills. The difference in effect sizes between on-site and virtual is statistically significant at a 5 percent level, for both outcomes. Regardless of-site coaching being about 23 percent more expensive than virtual coaching, it is more cost-effective.

Second, both programs were more effective at improving oral language proficiency than improving the somewhat more advanced skill of reading. Oral language proficiency is a precondition for reading in a new language and is the main focus of the English First Additional Language curriculum in grades 1 and 2. Quantile regressions reveal that the impact of on-site coaching on English reading was driven by the top half of the performance distribution. This may reflect the substantial gaps that exist in countries like South Africa between curriculum expectations and existing levels of learning. This is also the most significant sub-group effect that was identified.

Third, analysis of program implementation fidelity and intermediate outcomes suggest that the main reason for virtual coaching being less effective than on-site coaching was due to the modality of coaching rather than because of the format of the lesson plans (electronic versus paper-based) or quality of implementation. The same service provider implemented both programs, and attendance at teacher training was consistently above 90 percent for both interventions across the three years. On-site coaching visits happened more or less at the planned frequency (3 per quarter). Compared to on-site coaching, engagements between teachers and the virtual coach were more dependent on teacher cooperation (e.g. submitting portfolios of work to the coach or asking questions about the teaching methodologies)— and this varied widely across teachers. The reliance on a degree of self-motivation was identified in the qualitative research component as a possible weakness in the virtual coaching intervention.

Self-reported use of lesson plans and reading materials was relatively high in both intervention groups, but fidelity in following the daily lesson plans is hard to observe. However, we do have rich tablet usage data for teachers in the virtual coaching program and this points to rather low and varying levels of curriculum coverage. We estimate, for example, that only 27 percent of teachers accessed more than 60 percent of slides in the daily lesson plans. Slide coverage was better earlier in the Term than towards the end of the term but was highest in the week of learner assessments. This suggests that teachers were able to access the slides on the tablets when they perceived it to be really important but may have struggled to keep up with the learning program. This pattern would suggest that the technology itself was not the main barrier to program implementation, but rather the motivation of teachers or their ability to keep pace with the curriculum.

Perhaps, therefore, the differences in coaching modality account for the different program effects. Questionnaires indicated that teachers in the on-site coaching intervention were more aware of the components of the program than those in the virtual coaching intervention, and were more likely to have been observed teaching or to have seen a coach modelling teaching practices. Lesson observations indicated that teachers in both intervention groups were more likely than teachers in control schools to implement a wider spectrum of core curriculum activities and more frequently, but activities requiring more individualized attention to learners and higher-order pedagogical skills, such as group-guided reading and independent reading, were better implemented by teachers who had received on-site coaching. This suggests that the direct observation and opportunities for feedback available to an on-site coach were ultimately critical to program success.

1. Background and Context

This report builds on the information provided in the Year 1 and Year 2 reports of the Second Early Grade Reading Study (EGRS II). Detailed information about the situation of reading in both the Home Language (HL), English as the First Additional Language (EFAL), as well as prior research was provided in these reports and will not be repeated here. The focus of the Year 3 report is on the 2019 data collection processes and the main results after three years of implementation. This report provides information about the implementation of the interventions in Year 3 and the analysis of the learning outcomes at the end of Year 3. Details of the study site, the school selection process and the evaluation design are contained in the Year 1 report. A comprehensive final report on the EGRS II will be made available following the multi-method data collection, analysis and interpretation at the end of Year 4 (2020).

As a summary, the EGRS II is a Randomised Control Trial (RCT) that evaluates two early grade reading interventions in 180 primary schools in two districts in the province of Mpumalanga, South Africa. The EGRS II was first implemented with Grade 1 teachers in 2017, in 2018 the interventions were targeted at the Grade 2 teachers and in 2019 with Grade 3 teachers. Across all three years of the implementation, the study focused on measuring the causal impact on learner reading performance and unpacking the change mechanisms of a structured pedagogic programme.

1.1. Intervention design

The EGRS II focused on the early learning of English as a second language (officially named English as First Additional Language, or EFAL, in the South African curriculum) by providing specific resources, training and on-going coaching to teachers. The interventions that were trialled were based on the official government curriculum, formally referred to as the National Statements Grades R – 12. As such the interventions were designed to improve and strengthen *teachers' enactment of the official curriculum*, and not to evaluate and comment on the curriculum.

Both interventions consisted of three components: (1) detailed lesson plans, (2) integrated learning and teaching support materials and (3) instructional coaching and training. The main difference between the two intervention models was in the delivery model of the lesson plans and the coaching support. In intervention 1, the teachers received a paper-based version of the lesson plans and benefit from regular on-site coaching with a specialised reading coach that visited the teachers in their classrooms and providing clustered needs-based workshops after school. In intervention 2, the teachers received a tablet with an electronic version of the lesson plans, including various audio-visual resources and are supported through an Information and Communication Technology (ICT) coaching model that included telephone calls and cell phone messaging. The electronic lesson plans were provided on an application which was specifically developed for the study. The application was available offline to ensure functionality without data and connectivity concerns. This application contained additional electronic resources such as short training videos, sound clips of the phonics sounds, songs and rhymes and examples of learners' work.

Teachers from both interventions received training at the start of each term. The first training session was residential training and entailed two days of training for intervention 1 and three days of training for intervention 2, with the additional day spent on orientating the teachers to the tablets. The remaining training sessions were one-day cluster training with smaller groups of teachers. The on-site coaches trained the teachers that they were coaching, but because there was only one virtual coach, additional trainers were utilised to assist with the training of the intervention 2 teachers. The trainers rotated so that once during the year, all of the teachers in this intervention would be trained by the virtual coach once. If teachers from either intervention group did not manage to attend the training session, the on-site coaches organised a catch-up session to make sure that the teachers have the new materials and understand the instructional practices which were covered during the training.

In intervention 1, teachers received visits from specialist reading coaches about once a month. During these visits, coaches modelled, supported and evaluated teachers' practices and monitored implementation fidelity. Coaching in intervention 2 involved a phone call to each teacher once every two-weeks, regular text messaging and the establishment of virtual communities of practice. The virtual reading coach used text messaging provide teachers with weekly teaching tips, answering questions on the lessons and running bi-weekly competitions to see evidence of teachers' enactment of the lesson plans.

Table 1: Comparison of intervention 1 and intervention 2

	Intervention 1	Intervention 2
Provision of lesson plans	Paper-based	Electronic On an application on a tablet
Provision of LTSM	Paper-based: - Big books - Posters - Flashcards - Writing frames	Paper-based: - Big books - Posters - Flashcards - Writing frames
Coaching	Coach visits the teacher in her classroom. Once every three weeks.	Coach contacts the teacher via telephone calls and instant messaging (WhatsApp). Once every two weeks.
Training	**Initial training:** 2-day block training **Quarterly training:** 1 day at the start of each term **Needs-based training:**	**Initial training:** 3-day block training **Quarterly training:** 1 day at the start of each term **Needs-based training:**

	As required	Bi-weekly competitions[1]
Core methodologies	Paper-based instructional manual	Application-based instructions, Includes videos, sounds clips and photos of example writing

The teachers in intervention 2 were also supplied with videos, to assist the virtual coach in 'modelling' lessons to the teachers. The video technology operated in three different ways. Firstly, there were videos demonstrating core methodologies which were pre-loaded onto the tablet. The teacher could access these through the tablet and these videos showed a teacher or reading coach demonstrating the methodology in an authentic (classroom) context. A majority of these videos were filmed in the EGRS II classrooms. Therefore, teachers see the methodologies enacted by teachers like themselves in classrooms that look similar to their own.

The second two types of videos were not pre-loaded onto the tablet, but rather, were utilised in an organic way within the WhatsApp group. For example, at the beginning of each week, the virtual coach films herself saying the phonic sounds and words for the coming week. This is useful because it reminds teachers of what they are meant to teach that week and because English phonic sounds can be particularly challenging. In addition, the virtual coach films videos of challenge areas for teachers based on her conversations with teachers. These videos were either filmed from her desk or a classroom.

Finally, videos were used in the bi-weekly competitions. In this case, teachers were asked to take pictures or film videos in their classrooms of their practices. This is an opportunity for the virtual coach to gain eyes into the classrooms she is supporting and to see how teachers are implanting the instructional practices and core methodologies. The video (or picture) submissions, in turn, were designed to help the coach to better assess the areas she should concentrate on.

1.2. Research questions

The EGRS II is designed as a randomized control trial which evaluates the difference between two different coaching models. Coaches mainly play two roles: (1) one of accountability, where they monitor teachers' implementation of the curriculum and (2) one of support, where they build a trust relationship with teachers and offer practical and targeted support on instructional practices. The enactment of these roles looks very different between the on-site coaches and the virtual coach and the study aims to evaluate whether both methods can be equally effective.

On-site coaches have the benefit of being in the classroom and being able to model new practices in the teachers' context and can thereby support the gradual development of new practice from

[1] The bi-weekly competitions provide a platform for teachers to showcase their 'good practice' in instructional techniques and how they build print rich classroom environments.

novice to expert. The presence of the in-class support allows for the development of professional accountability in an environment of trust, where the coach monitors and evaluates the teachers' teaching practices to encourage more productive teaching practices. The on-going support from the coach also encourages the teacher to keep up with the increased pace of the scripted lesson plans throughout the year. Through the building of a trust relationship, coaches can also support teachers with the emotional labour, i.e. stress, insecurity and anxiety associated with developing a new professional practice mid-career.

The virtual coach does not have the benefit of being in the classroom and depends on new forms of support and guidance on teaching strategies through a range of materials, teaching guides, videos and interactive support platforms that are available at all times to the teacher. These resources are intended to encourage more productive teaching practices among teachers. The virtual coach builds a relationship with the teachers through phone calls and cellphone messaging and is available anytime during the day to support and assist teachers with questions that they may have. Therefore, while teachers knew the virtual coach, their in-person contact with her was limited to twice during the year. Neither of these meetings was done in classrooms, but rather at a neutral training location. The on-going support from the virtual coach intends to encourage the teacher to keep up with the increased pace of the scripted lesson plans throughout the year.

The virtual coach faced three challenges that the on-site coach did not have. Firstly, the virtual coach did not make in-person classroom visits but communicated through phone calls and text messages. For teachers who might not be interested in implementing new practices or engaging with their coach, these modes of communication are relatively easy to ignore. Secondly, the monitoring of teachers' implementation of the curriculum was dependent on teachers' responses and could not be verified through observations, and thirdly, building a trust relationship is more difficult when done virtually relative to having face-to-face conversations.

To mediate the first challenge, the virtual coach would try to engage with the School Management Team at the school to see whether they could encourage the teachers to utilise the support from the virtual coach. The second challenge was addressed by the introduction of small competitions around specific themes where teachers were asked to submit photos and videos via WhatsApp. These photos and videos allowed the virtual coach to get a better sense of how teachers are implementing the core methodologies and use these to inform the type of support offered to teachers. The final challenge was much harder to mediate and the virtual coach only had the phone calls and WhatsApp conversations through which to build this relationship.

Our main research questions are therefore:

1. Did on-site coaching improve learning outcomes in EFAL?
2. Did virtual coaching improve learning outcomes in EFAL?
3. Did the impact on reading proficiency differ between the two coaching models?
4. Which model is the most cost-effective?

2. Implementation Fidelity

Before we evaluate the impact of the interventions, it is useful to first understand whether the interventions were implemented as intended. This section draws on the monitoring and evaluation information from the implementing service providers and focusses on the extent to which teachers attended the training sessions, whether the coaches managed to visit teachers at the intended frequency and whether teachers used the electronic lesson plans provided.

Table 2 reports the teacher attendance rates at the four training sessions that were held at the start of each term. Overall, the attendance rates were very high, but in the case where teachers did not manage to attend the training session, the coaches organised a catch-up session to make sure that the teachers have the new materials and understand the methodology that was focussed on. School management team (SMT) members were also invited to attend the training sessions so that they are aware of the support that they can provide to their Grade 3 teachers. Their attendance at the training was not compulsory and therefore we see much lower attendance rates than for the teachers.

It is interesting to note that the attendance of SMTs from the virtual coaching schools was much lower than the SMTs from the on-site coaching schools. The project management team noted this at the end of the term one training, but despite the virtual coach's increased efforts each term to encourage SMTs to attend the training, their attendance rates dropped lower. One of the reasons could be that the on-site coach had a better opportunity of building a relationship with the SMT when visiting the schools, whereas the virtual coach remains rather abstract, regardless of efforts made to build a relationship via phone calls and text messages.

Table 2: Implementation data - teacher attendance at training

		Total no. teachers	No. teachers trained	No. teachers resources	No. of SMTs at training
TERM 1	On-site coaching	86	83 (97%)	86 (100%)	38 (76%)
	Virtual coaching	85	84 (98%)	85 (100%)	31 (63%)
TERM 2	On-site coaching	86	85 (99%)	86 (100%)	85 (99%)
	Virtual coaching	83	83 (100%)	83 (100%)	25 (51%)
TERM 3	On-site coaching	86	85 (99%)	86 (100%)	36 (72%)
	Virtual coaching	82	82 (100%)	82 (100%)	19 (39%)
TERM 4	On-site coaching	86	79 (92%)	86 (100%)	32 (64%)
	Virtual coaching	82	80 (98%)	82 (100%)	14 (29%)

A second aspect to consider is whether the on-site coaches managed to visit the teachers as intended. Table 3 shows the number of teachers each coach was supporting[2], the number of planned visits to each teacher, the number of actual visits they made to a teacher and finally, the number of needs-based workshops they organised with clusters of teachers. The evaluating team only had access to data aggregated by coach for the first three terms, but on average the coaches managed to do 3.29 visits per teacher in term 4 (i.e. fully achieving their target).

Most of the coaches either reached their target of 3 visits per teacher per term or even overreached by doing 4 visits per term. In term 3, however, the one education circuit experienced rather severe teacher industrial action due to the province not having paid some teachers. During this time minimal teaching was happening (often only until 10:00 in the morning, so that learners could still benefit from the school feeding scheme), and coaches were not allowed at the schools. The teachers in the virtual coach intervention who were affected by the strike also refused to take calls from the virtual coach during this time. A robustness check will be done in section 5.4 to evaluate the effect of the strike on the overall results.

Table 3: Implementation data - coaching visits

	No. schools	No. teachers	Planned visits	Term 1		Term 2		Term 3	
				Visits	*Workshops*	*Visits*	*Workshops*	*Visits*	*Workshops*
Coach 1	10	10	3 per term	4	5	4	7	3	5
Coach 2	14	31	3 per term	4	23	3	22	2	8
Coach 3	10	16	3 per term	4	30	4	6	3	3
Coach 4	16	29	3 per term	3	28	3	15	2	4

Each intervention 2 teacher received on-going messages reminding them of the work that will be covered each week, as well as up to four phone calls from the virtual coach each term. The virtual coach had an initial phone call each term to all teachers during which she tried to determine how much additional support the teacher required. Based on teachers' responses, the virtual coach would do follow-up calls to the teachers who still required additional support.

One of the challenges that the virtual coach faced was observing the quality of the teacher's instructional practices. To try to get a better sense of how teachers are implementing the core methodologies, the virtual coach introduced small competitions around specific themes. For instance, in the term that she focussed on phonics, the virtual coach would ask all teachers to submit a photo via WhatsApp of one of the phonics activities. She then choose the best teacher in each of the teacher groups who wins a small amount of airtime. The competitions allowed the virtual coach to see what teachers consider their best practice and will be able to start a conversation with teachers based on their submissions. It also helped teachers to see what other

[2] The variance in the number of teacher is a result of the geographical distances between schools. Coach 2 and 4 were supporting schools in a high-density area and could therefore visit more than one teacher on some days.

teachers in their surrounding area are doing, thereby fulfilling the role of a virtual community of practice. On average, of the teachers who participated at least once, teachers participated in 2.5 of the 4 competitions. However, from table 4 it is also clear that just less than half of the teachers had a very limited level of engagement.

Table 4: Participation in virtual coach competitions

No. competitions	No. teachers	% of teachers
0	18	21.69
1	18	21.69
2	16	19.28
3	12	14.46
4	19	22.89

3. Year 3 data collection

Four different evaluation activities were conducted at the end of the third year of implementation, each aimed at providing a different perspective of whether the interventions were successful and the mechanisms which contributed to the success.

1. The first activity was the main data collection which entailed assessing the same sample of learners that were assessed in the first three waves of data collection.
2. The second activity entailed retesting a sub-sample of the learners who were assessed in the main data collection activity, as a fieldworker quality check. The same learners were also assessed on a more extensive vocabulary assessment in both the HL and in EFAL. The EFAL vocabulary assessment aimed to better understand the impact of the interventions on EFAL vocabulary development, whereas the HL vocabulary assessment was included to evaluate whether there may have been any spillover effects on HL vocabulary development.
3. The third activity was a classroom observation study that had well-trained researchers observe the HL and EFAL lessons of 53 schools in the sample. The purpose of the classroom observation study was to evaluate whether the interventions have led to any changed instructional practices.
4. The final activity was a set of case studies conducted to further interrogate the impact of the interventions from a qualitative perspective.

This report will provide a comprehensive description of the evaluation results from the main data collection and will integrate the findings from the other evaluation activities to provide additional insights in section 5.4, 5.5 and 5.6. However, each of the other evaluation activities will have their separate reports as well.

3.1. Grade 3 Instruments

During the wave 4 main data collection, both an oral and a written assessment were administered to the learners in the sample. Furthermore, interviews were conducted with the Principal and the teachers to gain some additional contextual information.

The learner assessments were designed to evaluate learners' language and literacy abilities at the end of each grade but were not designed to necessarily benchmark learner performance against curriculum requirements. Given this focus, care was taken to minimize a floor effect. All tests are designed to be orally administered by the fieldworkers and to be captured electronically on the Tangerine software. As was the case with the baseline, Year 1 and Year 2 testing, to test the targeted 20 learners within one school day, the tests were designed to take no longer than 15 minutes to administer. The evaluation team piloted the instruments in five schools two months before the main data collection, after which the required amendments were made.

Table 5 below shows the various assessment tasks that have been included across the four waves of data collection. The learner assessment at the end of Grade 3 included an oral assessment that included seven tasks assessing HL and EFAL oral and reading proficiency. A further written assessment was conducted with the learners to assess their written comprehension abilities in both HL and EFAL, as well as a very short mathematics task.

Table 5: Learner assessment tasks across the various waves of data collection

	Construct	**Baseline**		**Year 1**		**Year 2**		**Year 3**	
		Start - Gr 1		End - Gr 1		End - Gr 2		End - Gr 3	
		HL	*EFAL*	*HL*	*EFAL*	*HL*	*EFAL*	*HL*	*EFAL*
Language Comp	Receptive Vocabulary		x		x		x		
	Expressive Vocabulary	x	x	x	x		x		x
	Listening Comprehension	x			x		x		x
Decoding	Phonological working memory	x							
	Phonological Awareness	x			x				
	Rapid Letter Naming					x		x	
	Letter-sound recognition	x		x		x		x	
	Word reading fluency	x		x	x		x		x
	Sentence reading fluency	x							
	Oral Reading Fluency (ORF)					x	x	x	x
	Reading Comprehension					x	x	x	x
	Written Comprehension							x	x
Spelling	Spelling of a CVC word				x				
	Writing two words						x		

Rapid letter naming (RAN) was included in both the year 2 and year 3 assessments to understand the proportion of learners who suffered severe reading difficulties. The purpose of the RAN is to measure the speed of lexical access and therefore gives an indication of learners' phonological processing skills. There is no consistent evidence internationally that RAN skills can be improved and the interventions did not specifically focus on improving learners' RAN skills. For this reason, we will not be using the RAN tasks to differentiate between the intervention groups.

One of the specific reasons for the inclusion of tasks in HL, i.e. Siswati and isiZulu and mathematics, is to examine possible spill-over and crowding-out effects. In this context, crowding-out would involve teachers using more time for EFAL than what the curriculum has earmarked for teaching literacy in the HL and/or mathematics.

Table 6 shows the correlation between the overall index scores between the assessments. From this, it is clear that the correlations between the later waves are slightly stronger than the correlations with the baseline assessment. This is mostly due to the difficulty of measuring literacy and language outcomes at the start of formal schooling. Unfortunately, this means that the baseline scores do not play a very strong role in controlling for any difference at the start of the study.

Table 6: Correlation between the assessments across the waves

	Wave 1	Wave 2	Wave 3	Wave 4 Oral Lang	Wave 4 Reading
Wave 1	1				
Wave 2	0.4102	1			
Wave 3	0.3999	0.7487	1		
Wave 4 EFAL Oral Language Proficiency	0.3859	0.5339	0.6548	1	
Wave 4 EFAL Reading Proficiency	0.3602	0.656	0.8366	0.6913	1

3.2. Data Collection

The wave 4 data collection was conducted by external service providers. Fifteen fieldwork teams were assigned to each assess 12 of the 180 sample schools. Although it was not possible to perfectly randomise fieldworker teams to school groups, the evaluation team worked with the service provider to ensure that each fieldworker team visits a good balance of schools from each of the intervention and control groups.

Fieldworkers were trained for five days of which the first two days focussed solely on the learner assessment tools and the third day of training was dedicated to in-school simulations. The final two days of data collection was dedicated to an assessor competency assessment, the contextual questionnaires and final logistical arrangements.

Fieldwork took place from 28 October to 15 November 2019 in 180 schools. At each school, fieldworkers collected data from the original sample of Grade 3 learners (some in other grades), Grade 3 teachers, and school principals. Fieldworkers were also tasked with assessing learners who had repeated Grade 1 or 2, in 2017 and/or 2018.

The main challenge experienced in the field were high rates of learners who had transferred to other schools since the baseline assessment in 2016.

Table 7: Instrument response rates

Research tool	Total Expected	Total Collected	Response Rate
Learner Oral Assessment	3327	2694	81%
Learner Written Assessment	2694	2661	99%
Teacher Questionnaire	NA	266	NA
Principal Questionnaire	180	180	100%
School Observation	180	179	99%

3.3. Re-test and Extended Vocabulary Assessment

In a subsample of learners from the main data collection, we administered the extended vocabulary assessment and also re-tested the learners on five of the sub-tasks. The re-test and extended vocabulary assessment was administered by a different set of fieldworkers but was administered on the same day with six learners per school from the main sample. The sample of learners was pre-selected by the evaluation team and included two learners at the top, middle and bottom of the performance distribution, based on the baseline letter recognition task. The purpose of the retest was to determine the extent of inter-rater reliability and the purpose of the extended vocabulary tasks was to get a more robust indication of learner vocabulary development. 315 learners from 60 schools participated in the vocabulary and re-test assessment.

For the vocabulary assessment, two picture vocabulary tests were developed (i.e. one in English and one in isiZulu/Siswati) to assess productive vocabulary. For each test, learners were presented with an image representing a concept (e.g. dog) and were asked to provide the word in the required language. Available vocabulary assessments are not necessarily appropriate for the South African context and are quite expensive to administer. The tests were therefore specifically developed for this purpose.

The two picture vocabulary tests were created for the study based on word frequency data in English and the medium of instruction (isiZulu or Siswati). Nouns, verbs, adjectives and prepositions from the British National Corpus (BNC)/Contemporary American English (COCA) and the Longman Spoken corpus which could be clearly represented in an image, were contextually appropriate and could be translated were selected, resulting in 214 words (items). These items were split into two tests of 132 items each (overlap of 49 items) and sorted by the English word frequency data i.e. higher frequency (better known) words were presented before lower frequency (less well known) words.

The tests were individually administered. A hardcopy file including one image per page was presented to participants. The HL test was administered before the English test. Before each test, learners received instructions in their HL and completed two practice items to ensure they understood the requirements of the task. Thereafter, prompts were given in the language of the test. Fieldworkers read the instruction from and scored on a tablet using Tangerine. Each test took a maximum of 15 minutes to administer, but due to the cut off rule, the total time varied per participant. The total score for each test was the sum of all the correct answers provided in the required language.

3.4. Lesson observation study and case study

In the same 60 schools in which the re-test and extended vocabulary assessments were administered, we conducted the HL and EFAL lesson observations. Due to protest action that was unrelated to the research study, we were unable to observe the lessons in two control schools, three intervention 1 schools and two intervention 2 schools. In addition to the lessons observed, we also conducted a more in-depth document review of learners' written exercises, as well as interviews with the teachers.

The case study used a unique case sampling methodology, in which unique cases provide for ways of developing or extending theories. In order to select cases (schools and teachers), a two-step process was followed. First, the researcher used empirical evidence from the Grade 1 and 2 end-of-year learner assessments to identify and select schools that performed above the mean in several key indicators (vocabulary, listening and speaking). Secondly, the virtual coach identified several schools where she felt that teachers were successfully implementing the tablet technology and utilizing the WhatsApp group and virtual coaching to drive their development. The schools that were both empirically above the mean indicators and were singled out by the coach were included in the study.

Following this selection process, both the virtual coach and the teachers were interviewed and observed. In total, sixteen teachers were interviewed and observed. This included eleven Grade 2 teachers in four different schools, and five Grade 3 teachers in three different schools were interviewed and observed. The Grade 2 teachers were interviews twice during the year. The goal of the first round of interviews was to gather information about the processes, and to generate hypotheses based on the information gathered. The second round of interviews asked more in-depth questions around the virtual coaching processes specifically. Finally, the Grade 3 teachers were interviewed once. These interviews were specifically intended to gather information around how teachers utilized different resources provided on the tablet.

3.5. Balance at baseline

As reported in the baseline report, the sample was balanced on the baseline assessment at the start of Grade 1. There is a slight imbalance on one of the sub-tasks, but since we are making 20 comparisons below, this is in line with what is expected. Moreover, the p-value of the F-test shows that we cannot reject the null for the joint significance across all the indicators. There is therefore no evidence of imbalance.

Table 8: Balance tests per task

	Control	**On-site**	**Virtual**	**On-site vs C**	**Virtual vs C**
	Mean/SE	*Mean/SE*	*Mean/SE*	*(1)-(2)*	*(1)-(3)*
Naming Animals in HL	7.155	7.310	7.501	-0.155	-0.346*
	[0.127]	[0.155]	[0.154]		
Word Recall	9.981	9.953	10.081	0.028	-0.099
	[0.084]	[0.093]	[0.092]		
Nonword Recall	4.208	4.179	4.237	0.029	-0.030
	[0.049]	[0.052]	[0.082]		
Phoneme Isolation	1.129	1.037	1.161	0.092	-0.032
	[0.087]	[0.092]	[0.107]		
Story Comprehension	2.179	2.154	2.263	0.025	-0.084
	[0.045]	[0.050]	[0.047]		
Letter Sounds Correct	6.978	6.784	7.019	0.194	-0.041
	[0.447]	[0.590]	[0.610]		
Words Read Correct	0.387	0.347	0.510	0.039	-0.123
	[0.096]	[0.103]	[0.148]		
Sentence Words Read Correct	0.051	0.027	0.034	0.024	0.018
	[0.012]	[0.011]	[0.012]		
Visual Perception	1.460	1.597	1.651	-0.137	-0.192
	[0.082]	[0.111]	[0.109]		
English Items	0.836	0.789	0.839	0.047	-0.003
	[0.044]	[0.063]	[0.045]		
N	1459	924	944		
Clusters	80	50	50		
F-test of joint significance (p-value)					0.782
F-test, number of observations					2383

Note. The value displayed for t-tests is the differences in the means across the groups. Standard errors are clustered at the school level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 8 below shows the balance of school-level characteristics. The school principals in the control schools are slightly older, and schools in the virtual coaching arm have a slightly larger problem of learner absence and have more dilapidated infrastructure. However, the F test for the join significance means that we cannot reject the null that these two samples are statistically equivalent.

Table 9: Balance on school characteristics

	Control	On-site	Virtual	On-site vs C	Virtual vs C
	Mean/SE	*Mean/SE*	*Mean/SE*	*(1)-(2)*	*(1)-(3)*
Principal Female	0.450	0.500	0.420	-0.050	0.030
	[0.056]	[0.071]	[0.071]		
Principal's age	52.550	50.120	51.160	2.430**	1.390
	[0.573]	[0.770]	[0.794]		
Grade 1 enrollment 2017	79.725	77.100	72.120	2.625	7.605
	[3.978]	[5.240]	[4.591]		
No. government teachers	1.950	1.960	1.860	-0.010	0.090
	[0.111]	[0.178]	[0.121]		
Vacancies of Grade 1 Educators	0.063	0.040	0.060	0.022	0.003
	[0.027]	[0.028]	[0.034]		
Problem - teacher absence	3.513	3.480	3.480	0.033	0.033
	[0.067]	[0.091]	[0.096]		

Problem - learner absence	2.975 [0.080]	2.900 [0.119]	3.180 [0.089]	0.075	-0.205*
Describe school maintenance	3.325 [0.090]	3.220 [0.125]	3.060 [0.141]	0.105	0.265*
N	80	50	50		
P-value				0.461	0.274
Number of observations				130	130

3.6. Attrition

During the Year 3 data collection, 2,684 of the 3,327 learners who were tested during the baseline data collection were re-tested and successfully matched to their baseline results. The overall attrition rate of 19% is slightly higher than what has been found in previous studies. When breaking down the attrition rate by intervention group, the differences are not statistically significant, but from figure 1 it is clear that the attrition rate of learners in the control schools (18%) was slightly lower than the attrition rate among learners in the virtual coaching and on-site coaching schools (each at about 20%). There was no imbalance on learner or school characteristics and the probability of attriting.

Table 10: Percentage of learners tested during baseline, Year 1, Year 2 and Year 3 data collection

	Intended sample	Baseline		Year 1 *Number re-tested & matched*		Year 2 *Number re-tested & matched*		Year 3 *Number re-tested & matched*	
		Tested Number	*Percentage Tested*	*#*	*%*	*#*	*%*	*#*	*%*
C	1,600	1,459	91%	1,347	92%	1,190	82%	1,193	82%
I 1	1,000	924	92%	820	89%	772	84%	735	80%
I 2	1,000	944	94%	873	92%	799	85%	756	80%
Total	**3,600**	**3,327**	**92%**	**3,040**	**91%**	**2,761**	**83%**	**2,684**	**81%**

Figure 1: Attrition rate by intervention group

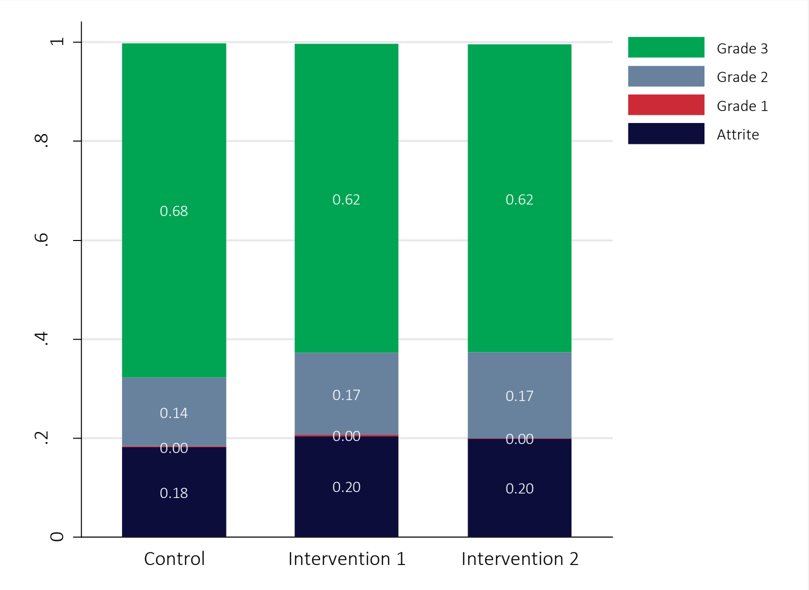


Table 8, however, shows that there may be some imbalance in the intervention groups in terms of learners who repeated a year or two. The first column suggests that learners in both the on-site and virtual coaching arms may have been more likely to repeat a year than learners in the control group. Further, younger learners, males and learners with lower scores at baseline seem to be more likely to have repeated a year. One of the robustness checks that we will do is to evaluate the impact of this imbalance on the final results.

Table 11: Learner progression in Wave 4 data collection

	(1)	(2)	(3)	(4)	(5)
	Delayed	Age	Gender	Zulu	BL Learning
Delayed		-0.144*	0.226***	0.040	-0.561***
		(0.086)	(0.040)	(0.048)	(0.075)
On-site coach	0.040*	0.016	0.029	-0.027	0.049
	(0.024)	(0.058)	(0.023)	(0.077)	(0.086)
Virtual coach	0.049*	0.056	0.004	-0.035	0.151**
	(0.027)	(0.060)	(0.028)	(0.078)	(0.074)
Delayed x T1		0.105	-0.104*	0.024	-0.125
		(0.114)	(0.055)	(0.066)	(0.129)
Delayed x T2		-0.110	-0.010	-0.066	-0.060
		(0.114)	(0.055)	(0.073)	(0.113)
Observations	2,684	2,684	2,684	2,684	2,684
R-squared	0.009	0.024	0.027	0.149	0.087
Mean attrition	0.174				

4. Task level learner assessment results

Table 11 provides information on the descriptive statistics of the assessment tasks administered at the end of Year 3. The scores in the table include the averages, an indication of the performance distribution, and the percentage of learners that scored zero on the task. The purpose of the table is to provide insights both to how learners on average performed relative to the maximum score in the task but also to provide a perspective of the relative distribution of scores. The table shows that there was a good distribution of scores in the Grade 3 assessment. The table therefore provides evidence that the assessment tasks provide sufficient information to differentiate learner performance across the sample distribution. It should be noted that unless otherwise specified, the descriptive statistics shown in this section only includes the scores of the learners who were still on-track (that is, in Grade 3 at the end of Year 3). In the appendix, the same tables are shown for the full sample of learners.

The zero scores indicate that the percentage of non-readers (i.e. the learners that could not read a single word correctly) was still remarkably high at the end of Grade 3. At the end of Grade 1, about 48% of learners could not read a single word correctly in HL. It is disconcerting that two years later, at the end of Grade 3, 18% of the Grade 3 learners still did not read a single word correctly in HL. The percentage of non-readers in EFAL was about the same (18%), even though the word length in the English language is shorter than in isiZulu and Siswati.

Table 12: Item descriptive statistics

	N	Mean	s.e.	p10	p25	p50	p75	p90	Min.	Max.	% zero score
Letter Naming TC	2148	22.2	8.5	11	17	22	28	34	0	36	1%
Letter Naming CPS	2148	1.1	0.4	0.6	0.9	1.1	1.4	1.8	0	3	1%
Letter Recognition	2148	47.7	21.3	18	33	48	62	76	0	110	1%
HL ORF at 60 seconds	2148	25.2	17.4	0	10	27	38	49	0	58	18%
HL ORF Comp	2148	2.6	1.9	0	1	3	4	5	0	5	24%
EFAL Word Recog	2148	27.7	21.5	0	7	27	43	56	0	99	15%
EFAL ORF at 60 secs	2148	33.5	30.9	0	2	29	56	75	0	126	19%
EFAL ORF Comp	2148	1.3	1.5	0	0	1	2	4	0	5	46%
EFAL Produc. Vocab	2148	3.6	1.7	1	2	4	5	6	0	6	5%
English Comp	2148	1.1	1.1	0	0	1	2	3	0	4	32%
HL Written Comp	2109	2.6	1.9	0	1	3	4	5	0	6	20%
EFAL Written Comp	2109	1.7	1.2	0	1	2	3	3	0	4	23%
Mathematics	2109	1.7	1.2	0	1	2	3	3	0	4	23%
EFAL Lang Prof. Index	2148	0.4	1.3	-1.1	-0.7	0.4	1.2	2.3	-1.8	3.4	
EFAL Read Prof. Index	2109	0.4	1.9	-1.9	-1.3	0.3	1.8	3.1	-2	5.8	
HL Read Prof. Index	2109	0.3	1.7	-2.4	-1.2	0.7	1.6	2.3	-3	4.2	

Notes: Sample only includes the Grade 3 learners. The statistics for the full sample is in table 1 in the Appendix. The construction of the Index is further explained in section 5.1

Table 12 shows the average scores for each of the sub-tasks by intervention group. Column (4) suggests that the learners in the on-site coaching intervention group performed better than the control group learners in HL letter recognition, EFAL word recognition, EFAL Oral Reading Fluency, reading comprehension, EFAL vocabulary and English listening comprehension. For learners in the virtual coaching group, the results were less significant, with some negative effects on the HL writing comprehension task.

Table 13: Tasks means in Wave 4, by intervention group

		On-site	Virtual	Control vs	Control vs
	Control	Coaching	Coaching	On-site	Virtual
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
Letter recognition	45.562	53.337	45.847	-7.774***	-0.284
	[1.351]	[1.709]	[1.530]		
HL ORF at 60 seconds	25.974	24.906	24.058	1.067	1.916
	[0.947]	[0.888]	[1.337]		
HL ORF Comprehension	2.681	2.681	2.472	0.001	0.209
	[0.094]	[0.114]	[0.130]		
EFAL Word Recognition	26.419	30.766	26.855	-4.346**	-0.436
	[1.052]	[1.329]	[1.590]		
EFAL ORF at 60 seconds	31.566	36.759	33.615	-5.192**	-2.048
	[1.480]	[1.676]	[2.284]		
EFAL ORF Comprehension	1.123	1.533	1.325	-0.410***	-0.203
	[0.070]	[0.104]	[0.114]		
EFAL Productive Vocabulary	3.354	3.856	3.622	-0.502***	-0.267*
	[0.087]	[0.091]	[0.128]		
English Comprehension	0.968	1.352	1.169	-0.385***	-0.201**
	[0.049]	[0.084]	[0.085]		
HL Written Comprehension	2.770	2.627	2.436	0.143	0.334**
	[0.099]	[0.129]	[0.132]		
EFAL Written Comprehension	1.616	1.756	1.615	-0.140	0.001
	[0.056]	[0.088]	[0.086]		
Mathematics	1.616	1.756	1.615	-0.140	0.001
	[0.056]	[0.088]	[0.086]		
EFAL Language Proficiency Index	0.168	0.643	0.418	-0.476***	-0.251**
	[0.061]	[0.086]	[0.102]		
EFAL Reading Proficiency Index	0.296	0.695	0.367	-0.399***	-0.071
	[0.091]	[0.121]	[0.146]		
HL Reading Proficiency Index	0.311	0.404	0.076	-0.094	0.235
	[0.096]	[0.114]	[0.132]		

The value displayed for t-tests is the difference in the means across the groups. Standard errors are clustered at the school level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Sample only includes the Grade 3 learners. The statistics for the full sample is in table 2 in the Appendix.

4.1. English Word Recognition and Oral Reading Fluency

Word recognition and oral reading fluency in EFAL are some of the central assessment tasks to evaluate the impact of the interventions on learners' ability to read in English. In the word recognition task, learners are given a chart of 104 words in English arranged from the simplest two-letter words to complex multi-syllable ten-letter words. Learners are asked to correctly name each of the words on the chart in sixty seconds. On average, the Grade 3 learners in the control group correctly identified 26 words in a minute. In contrast, learners in the on-site coaching group correctly identified 31 words, whereas learners in the virtual coaching group performed very similar to learners to the control group (27 words correctly). The proportion of learners in each group that could not correctly identify a single word was fairly consistent among the groups.

Table 14: Comparing EFAL word recognition across the waves of data collection

	End of Grade 1		End of Grade 2		End of Grade 3	
	Decodable words	Sight words	Decodable words	Sight words	Decodable & Sight Combined	% zero scores
Control	5	5.3	18.5	18.0	26.4	15%
On-site coaching	5.3	5.5	20.0	19.7	30.8	14%
Virtual coaching	4.6	4.7	18.0	17.9	26.9	17%

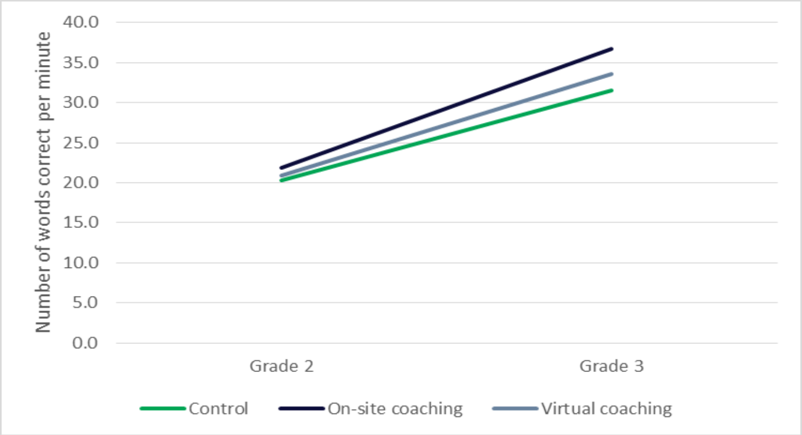
Notes: Sample excludes all repeaters. The statistics for the full sample is in table 3 in the Appendix.

In the previous round of testing, learners were assessed on decodable and sight English words separately. At the end of Grade 1, learners in all groups were only able to correctly read five words on average. At the end of Grade 2, that figure had increased to around 17 words per minute for both decodable and sight words but with little difference between the three groups. At the end of Grade 3, while all three groups of learners could read substantially more words correctly per minute, the learners in on-site coaching group were ahead of their counterparts in the number of words they could read correctly from a list.

Possibly the most important task within the Wave 4 learner assessment that could contribute to answering the main study questions is the assessment of learners' English oral reading fluency (ORF). Although there remains some question about the value of ORF and specific attained benchmarks (Kim et al 2010) as a predictor of reading difficulties, there is a growing consensus about the validity and reliability of ORF as a key measure of reading. It is now widely accepted that ability to read connected texts rapidly, accurately and with expression, is a critical competency required for successful reading for understanding.

The pattern that was seen in English word identification is equally clear in English oral reading fluency, with learners in the on-site coaching intervention group performing better than their peers in both the virtual coaching and control groups. At the end of Grade 2, there was very little difference in the reading fluency between the different intervention groups in EFAL (figure 2). The EFAL curriculum introduces an increased focus on reading in Grade 3 as can be seen in the increased reading fluency in the control group. It is encouraging to note that learners in both intervention groups saw a slightly higher rate of increase in their EFAL reading fluency than the control group.

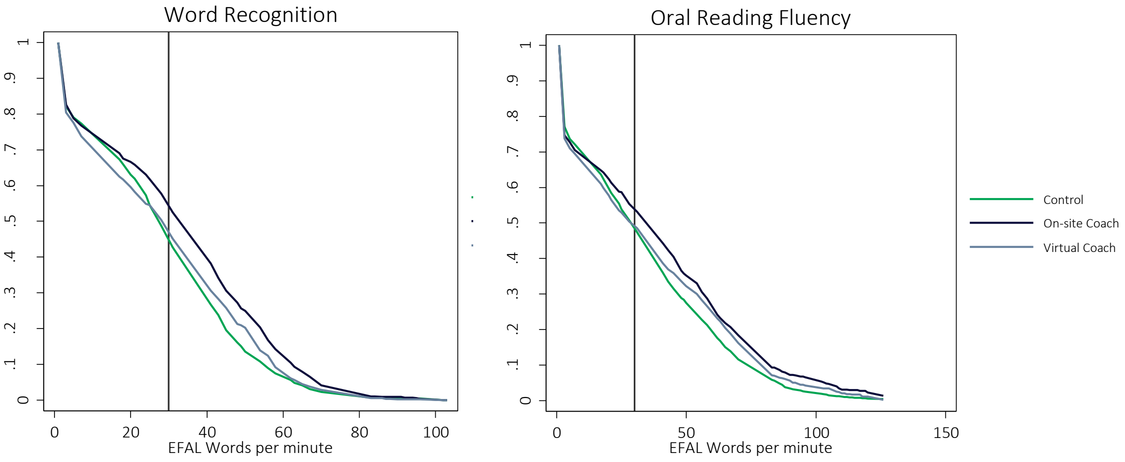
Figure 2: Improvements in oral reading fluency between Grade 2 and Grade 3



Notes: Sample excludes all repeaters. The statistics for the full sample is in figure 1 in the Appendix.

Figure 3 shows the performance distribution of learners in the three different groups for both word recognition and oral language fluency. Learners in the on-site coaching intervention out-perform learners in the virtual coaching and control groups across the performance distribution. The difference between the virtual coaching and the control groups are less clear, but it seems as if the virtual coaching group may perform slightly worse than the control group below the threshold of 30 words correct per minute, but that the picture changes thereafter with learners in the virtual coaching groups performing slightly better.

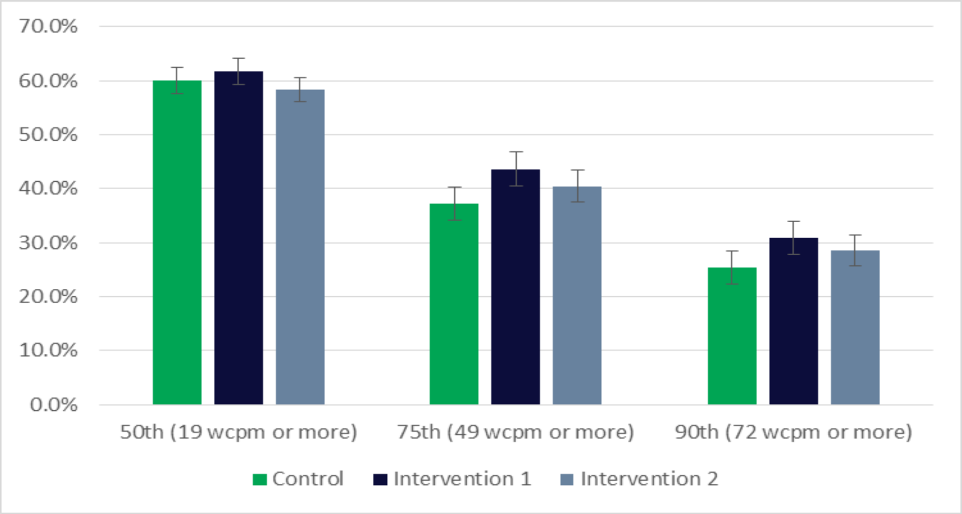
Figure 3: Performance distribution for EFAL word recognition and ORF



Notes: Sample excludes all repeaters. The statistics for the full sample is in figure 2 in the Appendix.

A useful gauge of the relative impact of the intervention is the proportion of the group that reach benchmark levels. Since there are no established reading benchmarks for EFAL validated and adopted in South Africa, we took the average words read correctly at the 50th, 75th and 90th percentiles respectively as possible thresholds. The figure below shows that there were a significantly higher proportion of learners in the on-site coaching intervention that reached the benchmarks at the 75th and 90th percentile, relative to learners in the control group. Although the proportion of learners in the virtual coaching intervention was also higher at these thresholds than the control group, the difference is not statistically significantly different.

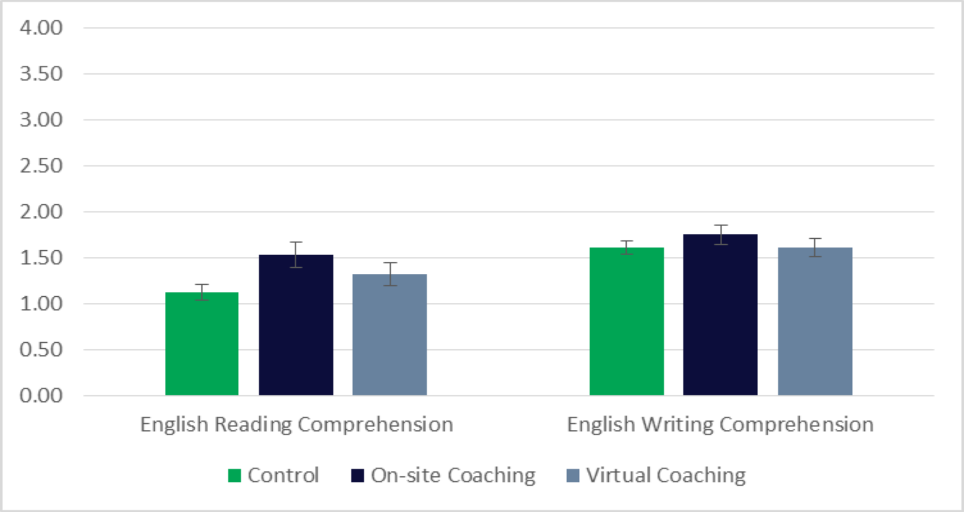
Figure 4: Proportion of learners that reached certain reading thresholds



4.2. English Oral and Written Reading Comprehension

One of the primary concerns is the extent to which the English reading interventions improved children's reading comprehension or understanding. To assess reading comprehension, we included an oral and written reading comprehension task. The oral reading comprehension task followed the oral reading fluency passage, whereas learners were given a passage to read independently and asked to answer the questions in a written format for the written comprehension task. There was no time limit for the responses to the five questions that followed the oral reading fluency passage, and learners were given three minutes to read the passage. Learners were subsequently only asked questions relevant to the section up to where they read. For the written comprehension task, learners were given a passage and four questions in a written format and were given eight minutes to complete.

Figure 5: English reading and writing comprehension

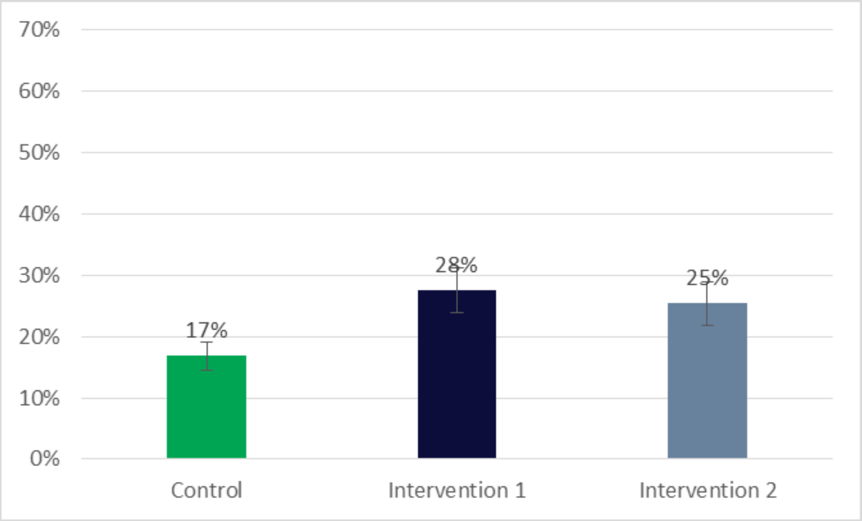


Notes: Sample excludes all repeaters. The graph for the full sample is in figure 3 in the Appendix.

Possibly the largest difference between the three groups was the learners' performance on the oral reading comprehension questions. When means scores include only Grade 3s, learners in the on-site group scored in the order of 40% higher than those in the control group. The magnitude of difference between the virtual group and the control was about half. In the English written comprehension tasks, the learners in the on-site coaching group seemed to have scored slightly higher than learners in the other two groups, however, this is not statistically significant.

If we are to assume that learners that can answer three or more of the oral comprehension questions correctly have a reasonable understanding of the passage, then we see that the learners in the on-site and virtual coaching are substantially ahead of learners in the control group.

Figure 6: Proportion of learners by intervention group that correctly answered three or more EFAL comprehension questions

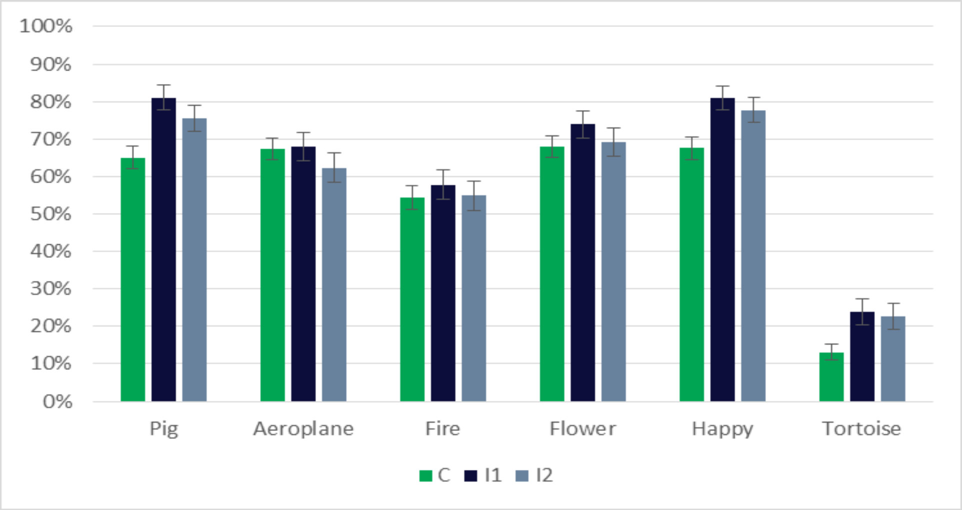


4.3. English Productive Vocabulary and Oral Comprehension

In addition to a parallel in-depth study of vocabulary, there were two tasks included in the end of Year 3 learner assessment to evaluate learners' English oral language proficiency. To measure expressive vocabulary, learners were shown pictures, and asked: "what do we call this in English?" And to measure listening comprehension learners needed to respond to a set of questions related to a short story that was read to the learner in English.

One of the striking findings in the Year 1 analysis was the similarity in the gains for learners in both intervention groups relative to the control on English productive and expressive vocabulary (Kotze et al, 2019). These results were central to the preliminary finding of the relative efficacy of both on-site and virtual coaching. At the end of the third year we see that learners in the on-site coaching intervention have improved marginally more in these learning outcomes than learners in the virtual coaching intervention.

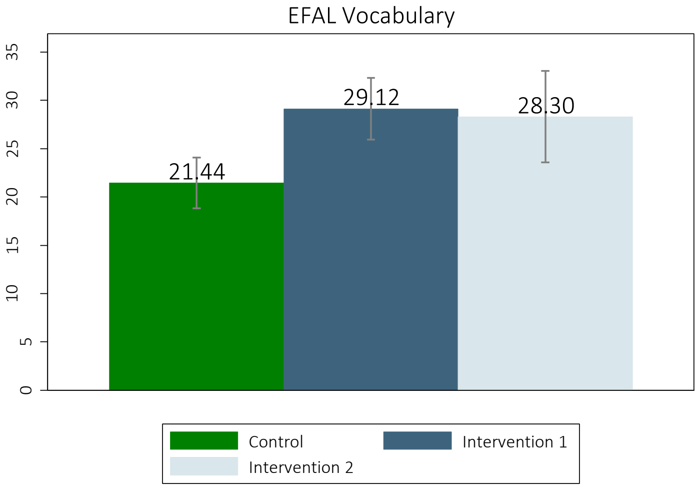
Figure 7: English Vocabulary in the main assessment



Notes: Sample excludes all repeaters. The graph for the full sample is in figure 4 in the Appendix.

A more in-depth vocabulary assessment was done with a subsample of learners. The purpose of this assessment was to gain a richer understanding of the English vocabulary that learners attained by the end of the foundation phase. On average, learners in both the on-site coaching and virtual coaching groups had a more extensive English vocabulary. Learners in the control group only managed to answer 21 of the 132 vocabulary questions correctly, whereas learners in the on-site coaching group attained 29 correct and in the virtual coaching group learners managed to answer on average 27 correctly.

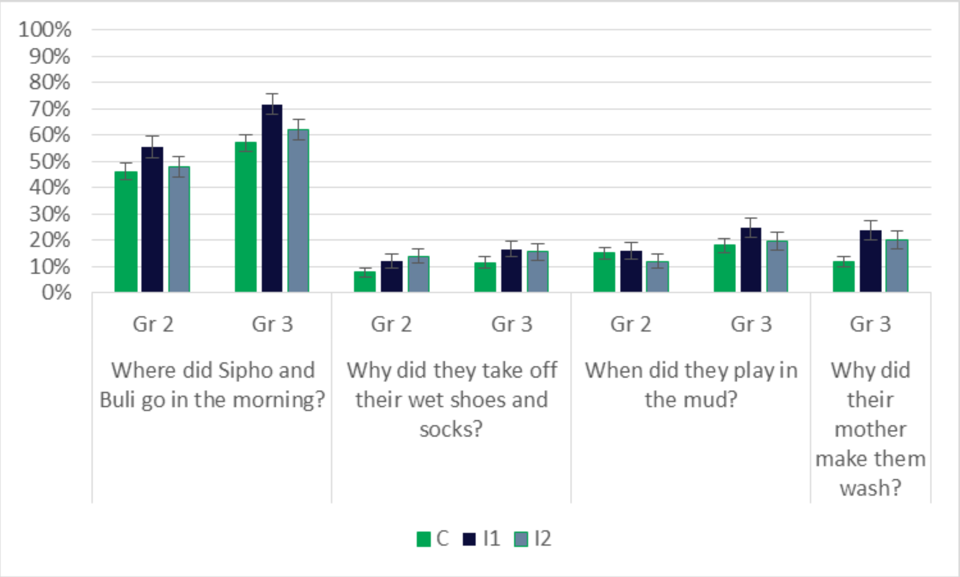
Figure 8: Extended vocabulary assessment



Notes: The extended vocabulary assessment was conducted on a sub-sample of learners.

The English listening comprehension was assessed by answers to questions after the reading of a short story to learners. The same passage and questions were administered to learners at the end of Grade 2. It is rather disconcerting that very few of the children overall could answer these questions. Nevertheless, the results of the intervention groups are marginally better than the results for the control group, with learners in the intervention groups outperforming the control learners on two of the three items. However, the vast majority of learners in the study were not able to answer questions based on an oral English story told to them. This suggests that while a larger group of learners were able to correctly identify objects in English, their mastery of even simple narrative interactive English is limited.

Figure 9: EFAL listening comprehension scores



Notes: Sample excludes all repeaters. The graph for the full sample is in figure 6 in the Appendix.

4.4. Letter Recognition

In the baseline assessment, the learners identified just fewer than seven letter sounds correct in their HL with about 19% of the sample scoring zero. Three years later, at the end of Grade 3, the average number of letter sounds correctly recognised increased to 44, and the percentage of learners not able to correctly identify any letter sounds was down to 2%. Although learners in the on-site coaching group were able to correctly identify five more letters than the comparable learners in the control group, there was no real difference between the control learners' performance and those in the virtual coaching group.

Table 15: Comparing HL letter-sound recognition

	Start of Grade 1	End of Grade 1	End of Grade 2	End of Grade 3
Control	7	17.7	38.9	45.6
On-site coaching	6.8	16.7	40.4	53.3
Virtual coaching	7	15.1	36.6	45.8

Notes: Sample excludes all repeaters. The statistics for the full sample is in table 4 in the Appendix.

The fact that all learners can correctly recognise letter sounds by the end of Grade 3 is to be expected as this is one of the most basic skills that should have been mastered within the first few months of Grade 1. It is interesting to note that through-out Grade 3 there continued to be a substantial improvement in this literacy indicator, which suggests that basic skills not acquired in the first two grades can be learnt in Grade 3 and continued levels of mastery happen into the final year of the Foundation Phase.

4.5. Home Language Oral Reading Fluency and comprehension

One of the core assumptions implicit in the assessment is that learners' mastery of reading in English requires that they have a strong foundation of reading in their home language. The first assessment task, i.e. HL letter recognition is a building block of reading in any language. Although not part of either the on-site or virtual coaching interventions, the assessment of HL literacy skills provide important insights into the foundations on which learners build their second language skills. The HL items also allow us to evaluate whether there are any positive or negative spill-over or crowding-out effects as a result of the interventions. Since the interventions focus on teaching EFAL, teachers may spend more time on teaching EFAL at the cost of teaching HL, which may have a detrimental effect on home HL. Alternatively, the skills teachers are teaching are transferable to teaching HL and may therefore enhance their teaching of HL and subsequently learners' HL outcomes.

Table 16: Comparing HL word reading

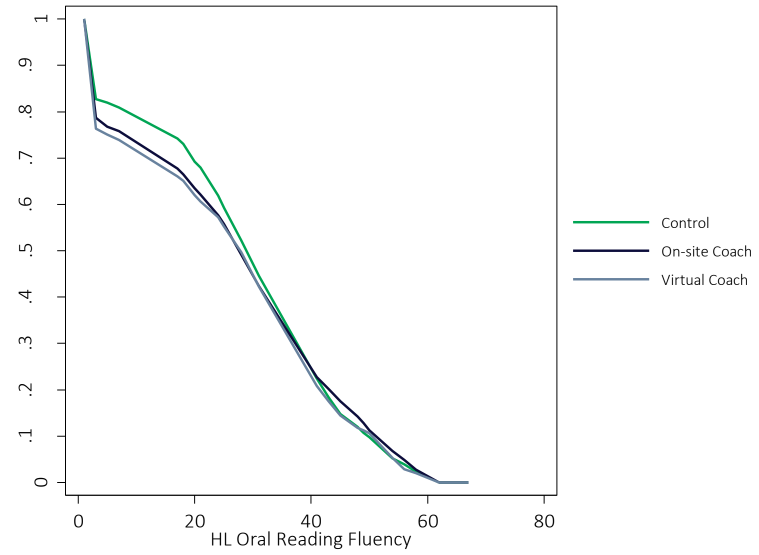
	Word Reading End of Grade 1		ORF End of Grade 2		ORF End of Grade 3		
	Average words	% of Zero Scores	Average words	% of Zero Scores	Average words	% of Zero Scores	Mean Comprehension
Control	5.5	45.7%	17.0	30.5%	26.0	16.0%	2.7
On-site coaching	4.7	49.1%	15.1	36.8%	24.9	19.6%	2.7
Virtual coaching	4.7	51.8%	14.5	40.6%	24.1	21.1%	2.5

Notes: Sample excludes all repeaters. The statistics for the full sample is in table 5 in the Appendix.

The first and most important insights of the ORF task in the HL is that 25% of the entire learner sample was not able to read a single word correctly in their HL at the end of the Foundation Phase. These learners had not made any progress in reading and comprehension in their HL from three years of schooling and are still not mastering skills and knowledge associated with the first year of formal schooling. While we are concerned about the high proportion of learners not meeting the minimum proficiency, we have major problem that a quarter of learners that reached Grade 3 had not gained even the most basic elements of literacy in their HL.

From figure 10 it is evident that at the bottom end of the performance distribution learners in the two intervention group are performing worse than their peers in the control schools. Further investigation is needed to explain why the HL oral reading fluency mean scores in the two interventions are below that of the control group and we will interrogate this more rigorously in section 5.

Figure 10: Distribution of HL oral reading fluency



Notes: Sample excludes all repeaters. The graph for the full sample is in figure 6 in the Appendix.

The average scores on the five comprehension questions that followed the HL oral reading fluency passage were very similar for the three groups (control, 2.7, on-site coaching group 2.7 and virtual coaching 2.5) and it does not appear that the interventions prejudiced the learners in these groups on their HL literacy development, particularly the core task associated with reading for meaning.

5. Main Results

5.1. Main Regression Findings

As specified in our pre-analysis plan, we have decided to evaluate the overall impact of the interventions using two indices that are based on the two language constructs that learners of a second language have to master in the Foundation Phase. The first construct is language proficiency as it relates to English vocabulary development and the second relates to decoding skills. In the first grade, learners are only taught language proficiency skills during the English lessons, whereas the decoding skills are already taught during the Home Language lessons in the first grade. Decoding skills are only introduced in the English lesson from the second half of the second grade and build on the skills that learners were already taught in their Home Language. By the third grade, both language proficiency and decoding skills are consolidated and learners should be able to read for meaning. The two primary outcomes that we will look at are therefore (1) English language proficiency and (2) English decoding.

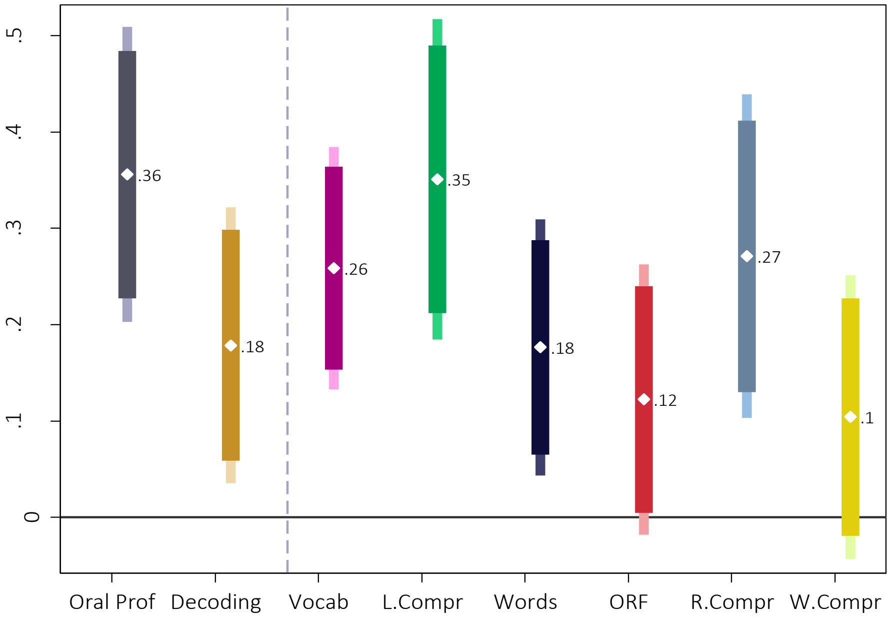
The indices are constructed using principal component analysis (PCA) which is a statistical method that combines the various subtasks under each construct into one single score. Intuitively, the PCA creates an index that is reflective of the most common underlying construct of the subtasks included in the index. The English language proficiency index is constructed using the English expressive vocabulary task and the English listening comprehension task, both tasks that assess English language use skills. The English decoding index is constructed using the English word recognition, English oral reading fluency, English reading comprehension and English written comprehension subtasks. Both indices were constructed using only the control group's scores as they serve as our reference group. The indices were then standardised on the control group mean and standard deviation so that the results can be interpreted in standard deviations relative to the control group (whose mean will now be zero).

Figure 11 shows the estimated impacts for the Grade 3 learners who were in the schools that received the on-site coaching intervention. The first two columns indicate the coefficients for the language proficiency and decoding indices respectively. The point estimate is indicated by the white diamond, the confidence interval at the 95% level is shown by the darker coloured bar and the confidence interval at the 90% level is shown by the lighter colour bar. The six bars after the dotted line shows the coefficients for the sub-tasks (which were also standardised around the control mean to allow for comparison with the indices).

The coefficients were derived from separate regressions run on each variable, controlling for the learners' scores on the baseline sub-tasks, learner gender, learner age, the education district, the quintile status of the school, the stratification dummies and fieldworker dummies. We decided which controls to include, based on the controls which explained the most variation in regressions run only on the control group. The regression table is shown in table 6 in the Appendix.

It is clear from figure 11 that the on-site coaching intervention had a positive and significant impact on both learners' language proficiency and decoding skills. The coefficients on the subtasks indicate which subtasks are driving these results. Learners in the on-site coaching intervention did significantly better than their control group peers in expressive vocabulary and listening comprehension. The improvements in the decoding skills were slightly less pronounced, with the improvement in oral reading fluency only being significant at the 90% level. In terms of raw scores, learners in the on-site coaching group read on average 3.6 more words correctly on average than their control group peers in the EFAL word recognition task, and 3.5 more words correctly in the ORF task.

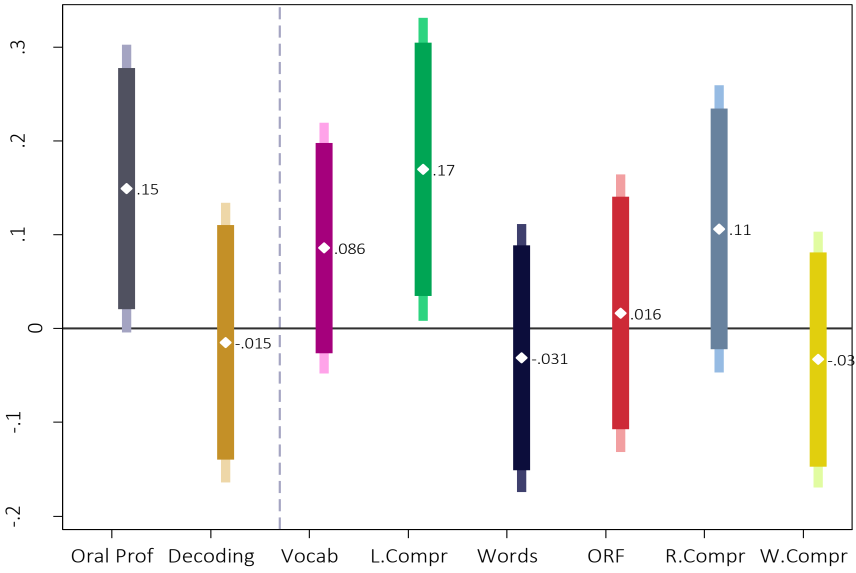
Figure 11: Learner performance on EFAL tasks – On-site coaching



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 7 in the Appendix.

Figure 12 shows that the impact of the virtual coaching intervention was much weaker than the on-site coaching intervention. The impact on language proficiency was less than half of the impact we saw in the on-site intervention and is only significant at the 90% level. The coefficient for the decoding skills is very close to zero, which means that there is no noticeable difference between the decoding skills of the learners in the control group and the learners in the virtual coaching group. In terms of sub-tasks, learners in the virtual coaching intervention only outperformed their control group peers in the listening comprehension task.

Figure 12: Learner performance on EFAL tasks – Virtual coaching

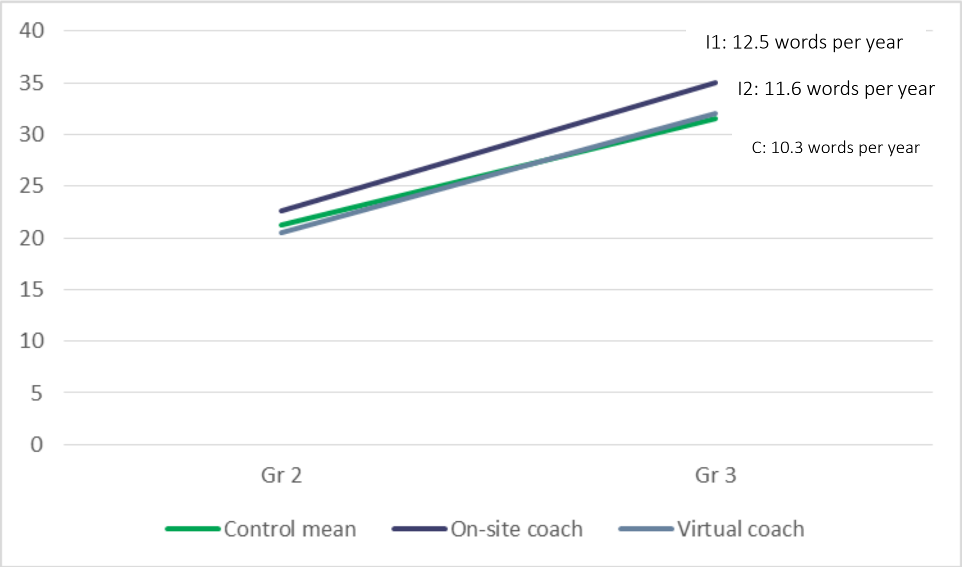


Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 8 in the Appendix.

Although standard deviation change is a useful measure of intervention impact, it is often difficult to understand the magnitude of the impact in terms of actual learning gains. To try to translate the gains in the intervention groups relative to the control group, we make use of the Oral Reading Fluency task. Albeit using different reading passages between the two grades, this task is relatively comparable between the two years.

Figure 13 is similar to figure 3, but is based on regressions run on the raw scores and therefore controls for the baseline differences between the intervention groups. Between the end of Grade 2 and Grade 3, the control groups read 10.3 more words correctly within a minute, the on-site coaching group read 12.5 more words and the virtual coaching group read 11.6 words more. The learning gains of the control group over the year can be interpreted as a year's worth of learning. Relative to the control group learning gains, we can then conclude that on oral reading fluency, learners in the on-site coaching intervention group learned 21% of a year more, and learners in the virtual coaching group 10% more (although the gains in the virtual coaching group are not statistically significant).

Figure 13: Learning gains in oral reading fluency between Gr 2 and Gr 3



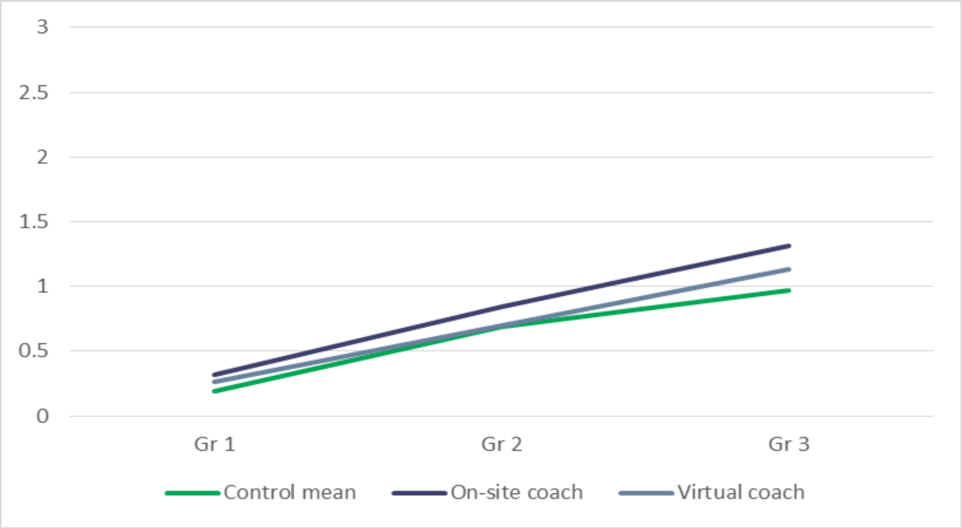
Notes: Sample excludes all repeaters. The graph for the full sample is in figure 9 in the Appendix.

Similarly, we can track the learning gains in the English listening comprehension task. The same English listening comprehension task was asked at the end of Grade 1, Grade 2 and Grade 3. The task entails the fieldworker reading a short English story (4 sentences long) and then asking the learners 3 questions.[3] Overall, the scores on the listening comprehension task were very low, but we nevertheless saw some significant gains in the intervention groups. Over the two years, we saw that learners in the control group score 0.7 points more on the task, the on-site coaching learners 0.9 points more and the virtual coaching learning 0.8 points more. This relates to learning gains of

[3] In Grade 3 the fieldworkers asked 4 questions, but only 4% of learners answered all 4 questions.

31% of a year's worth of learning for the on-site coaching learners and 10% of a year's worth of learning for the virtual coaching learners.

Figure 14: Learning gains in English listening comprehension between Gr 2 and Gr 3



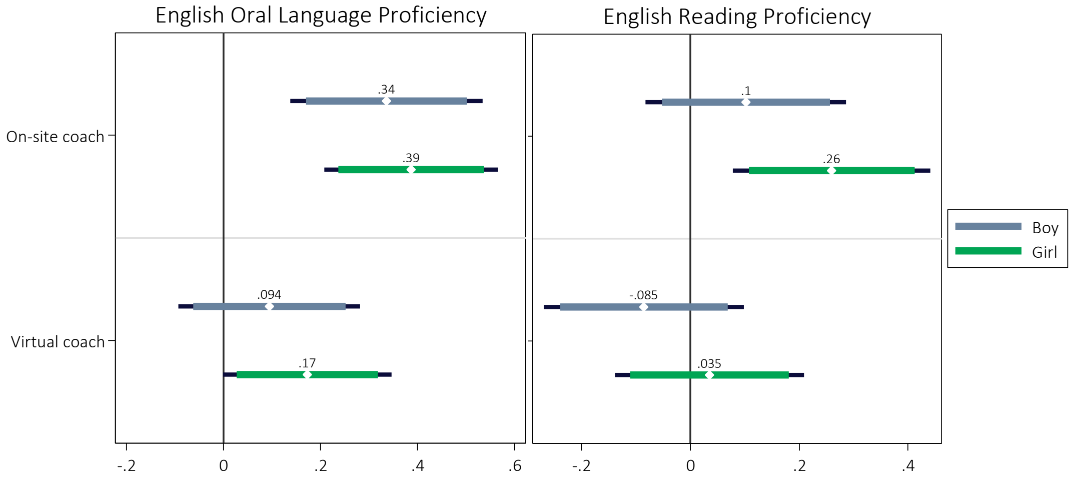
Notes: Sample excludes all repeaters. The graph for the full sample is in figure 10 in the Appendix.

5.2. Sub-Group Analysis

In our pre-analysis plan, we specified that we will evaluate whether the interventions impacted learners differently based on four different characteristics, namely the districts where the programme is implemented, learner gender and learner ability and the home language of the learner.

In South Africa, we find that girls outperform boys in reading outcomes from a very young age. The baseline assessment showed that girl performed better than the boys in tasks such as non-word recall, letter recognition and phoneme isolation. Boys may benefit more from the interventions as a result of more structure imposed in the classroom through the programme. This is something we observed in the first two years of the first Early Grade Reading Study, where the interventions seemed to have helped boys catch-up to girls. However, this trend was not sustained one year after the interventions were concluded.

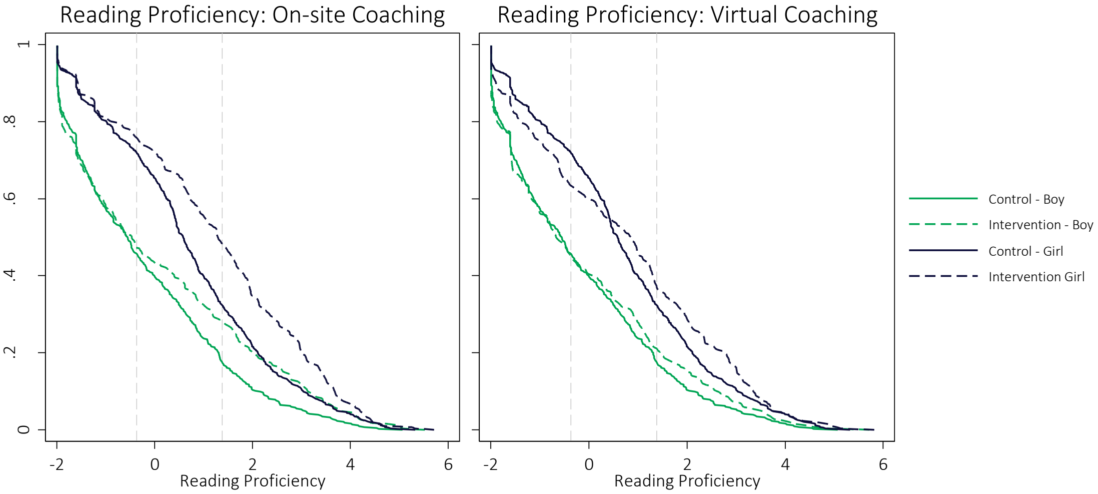
Figure 15: Girls performing better than boys in reading proficiency



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 11 in the Appendix.

Figure 15 suggests that girls may have benefitted slightly more than boys from the on-site coaching intervention in their reading proficiency skills, but that this additional benefit is not statistically significant. Interrogating this further, figure 16 firstly shows that girls performed better than boys in reading proficiency across the distribution. What is striking about figure 16, is that the girls in the control group still outperform the boys in the intervention group. Both girls and boys in the top half of the performance distribution benefitted more from the on-site coaching intervention, but there is no noteworthy difference in the virtual coaching intervention.

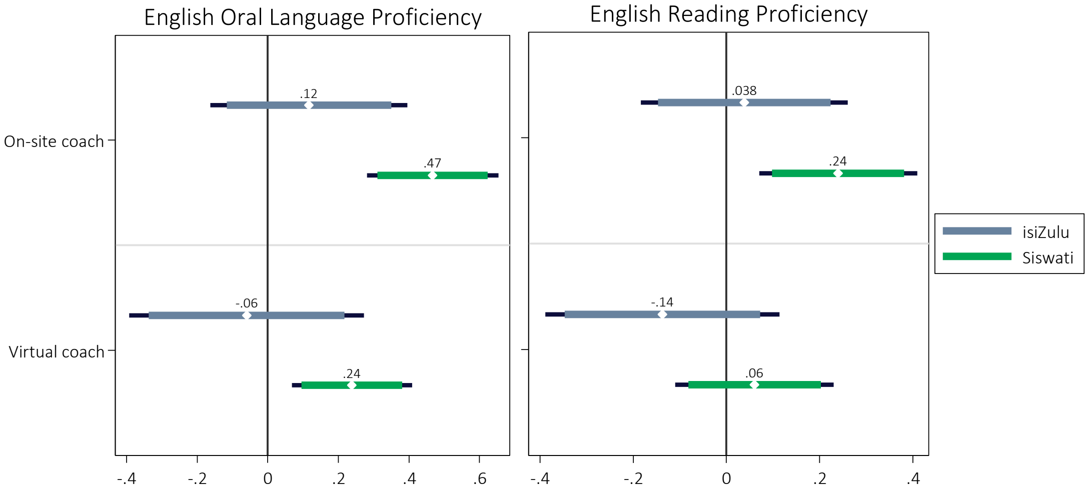
Figure 16: Girls benefitting more from the on-site coaching intervention in reading proficiency



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 12 in the Appendix.

Results after year 1 and year 2 of the interventions suggested that learners in the schools with Siswati as the language of learning and teaching may be benefitting more from the interventions than learners in the isiZulu schools. Given that the interventions were conducted in English, targetting EFAL, we did not expect there to be a difference based on the language of learning and teaching in the schools and therefore we did not stratify the sample based on the language of learning and teaching in schools. 71% of the learners in our sample are in schools with Siswati as the Language of learning and teaching (LoLT), whereas the other 29% are in isiZulu schools. In the Year 2 report, we did an extensive interrogation of the factors that may be driving the differential effects and saw that the isiZulu schools may be located in regions that may struggle to attract new teachers and in regions where the teachers are more likely to live further away from the school. However, none of the factors conclusively accounted for the differences.

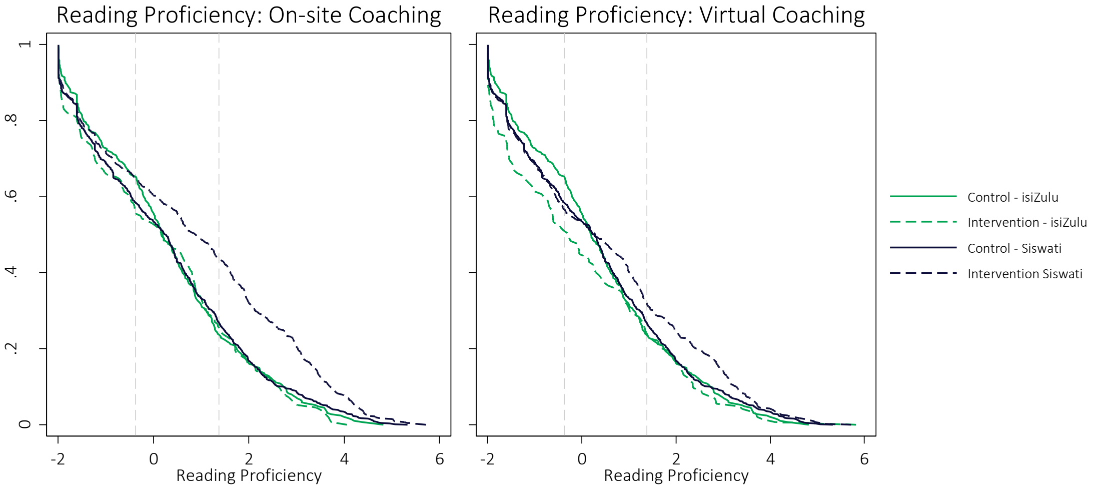
Figure 17: Learners in Siswati schools seems to be benefitting more from both interventions



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 13 in the Appendix.

In year 3 we again see some differential impacts in the Siswati schools, but they are much less pronounced than at the end of year 2. Figure 17 shows that learners in the Siswati schools may have benefitted slightly more from the interventions in both oral language proficiency and reading proficiency, but the difference is only statistically significant at a 90% level for on-site coaching in oral language proficiency. Figure 18 considers the differential impact across the performance distribution on oral reading fluency and suggest that learners in the upper part of the performance distribution may have seen higher gains from on-site coaching.

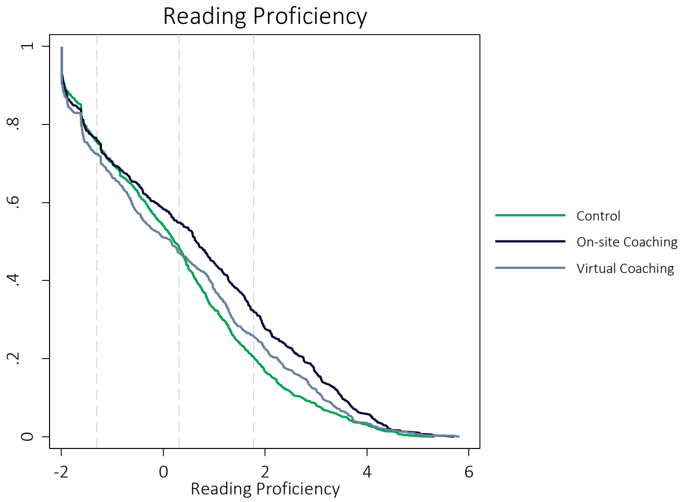
Figure 18: Learners in Siswati schools benefitting more from the on-site coaching intervention



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 14 in the Appendix.

No noteworthy differential impacts were observed when disaggregated by learner baseline ability or by the educational district. Evaluating any differential impact based on baseline ability is, however, problematic because the correlation between the wave 4 and wave 1 scores are particularly weak. Another way of evaluating whether stronger or weaker learners may be benefitting more is by looking at the learner performance across the learning distribution. The distribution of the reading proficiency index is shown for each intervention group in Figure 19. For each ability level across the performance distribution, the line shows the percentage of learners from each intervention group that could reach that ability level. For example, the first dotted line indicates the level at which learners at the 25th percentile performed and we see that at least 75% of learners reached at least this level. Similarly, the second dotted line shows the level at the 75th percentile and we see that 30% of learners in the on-site coaching intervention, 23% of learners in the virtual coaching group and 20% of learners in the control group managed to reach at least this level. This graph therefore shows us that learners in the top half of the performance distribution (above the 50th percentile) seem to have benefitted more from the on-site coaching intervention than learners in the bottom half of the distribution. Again, no real difference is seen across the performance distribution between the virtual coaching group and the control group.

Figure 19: Difference between the intervention groups across the performance distribution.



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 15 in the Appendix.

Another way of trying to understand whether weaker or stronger performing learners may be benefitting more from the interventions would have been to consider learner's HL proficiency. The South Africa curriculum is developed based on the additive bilingual approach which is that learners build their additional language skills of the base of their HL skills. This suggests that learners who have a stronger HL base may be able to benefit more from an EFAL intervention. However, section 5.3 below shows that the interventions did influence the HL performance, which renders this option also impossible.

5.3. Secondary outcomes

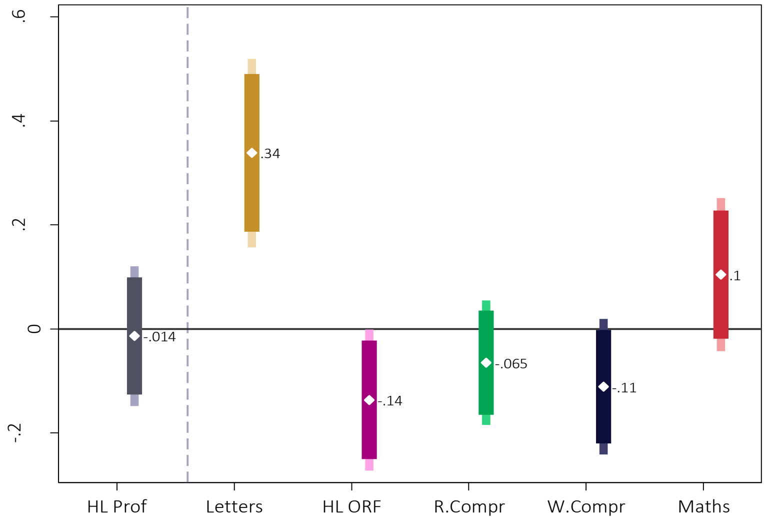
We will consider two separate sets of secondary outcomes. Firstly we will see whether the English interventions had any crowding-out or spillover effects on the other two subject areas (HL and Mathematics). Secondly, we will examine the impact of the interventions on a series of intermediate outcomes to see which mechanisms may be driving the results in the primary outcomes.

5.3.1. Crowding out and spillovers

The South African curriculum specifies that four subjects should be taught during the foundation phase: HL, EFAL, Mathematics and Life Skills. The interventions only supported teachers with the teaching of EFAL, but there are two ways in which the teaching of the other subjects could have been influenced by the interventions: teachers could either dedicated more time and effort to the teaching of EFAL, at the cost of time and effort spent on teaching the other subjects (crowding-out), or teachers could have applied the more effective teaching methodologies that they were taught in the EFAL intervention to the teaching of the other subjects (positive spillover). To test which effect was more dominant, learners were assessed on a few HL reading tasks, as well as a very short mathematics task.

The HL tasks included letter recognition, oral reading fluency, five comprehension questions based on the oral reading fluency passage and a written comprehension assessment. Figure 20 below shows that the coaching intervention had a strong positive effect on letter recognition, but a negative impact on HL oral reading fluency. Although both comprehension tasks have a negative coefficient, these are not statistically significant.

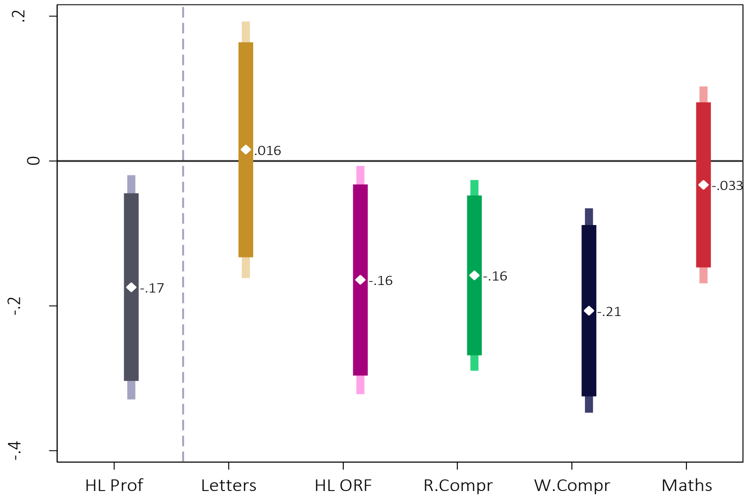
Figure 20: Effect of on-site coaching on HL and Maths



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 16 in the Appendix.

The virtual coaching intervention seemed to have had a more negative impact on HL outcomes. Learners in the virtual coaching intervention performed worse than their control group peers on HL oral reading fluency, reading comprehension and written comprehension. The negative effect on the HL items are rather curious and can point to possible crowding-out effects. However, there was no noteworthy impact on mathematics which would suggest that the trade-off for time only occurred between the two language subjects.

Figure 21: Effect of virtual coaching on HL and Maths



Notes: Sample excludes all repeaters. The graphs for the full sample is in figure 17 in the Appendix.

These results leave us with questions on why we are seeing the negative spillover effects on the HL subtasks and why this effect is more pronounced for virtual coaching. These questions will be investigated more thoroughly in section 6 of the report.

5.3.2. Mechanism driving change

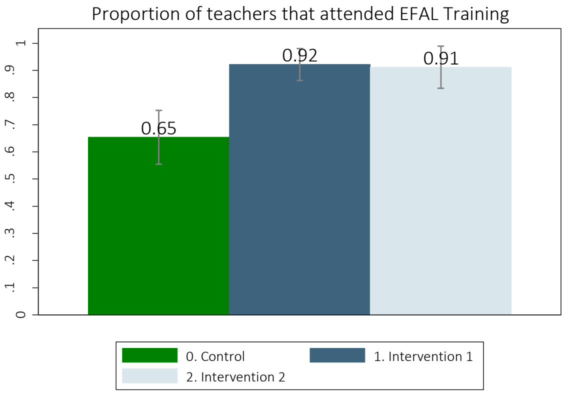
This section will aim to see whether there is any evidence of the theory of change being realised. The interventions aim to affect instructional practice change amongst teachers at a large scale, in line with the curriculum and methodologies in which teachers were trained during the teacher training at the start of the programme. The scripted lesson plans provide a mechanism to prompt the enactment of the behaviour change, whereas the coaching serves as an additional mechanism to encourage fidelity to the programme. We will therefore look at implementation fidelity from the teachers' perspective (rather than service provider reports), teachers' reports of instructional practices change, observed teacher instructional practices change, change in teachers' skill acquisition and evidence of any change in-school support by SMT members.

Implementation fidelity:

To evaluate implementation fidelity, we asked teachers whether they attended EFAL training in the year, whether they received support from their coaches, whether they had access to graded readers (these were supplied to all intervention school teachers) and how much time they spent on teaching EFAL. These questions will first be looked at descriptively, but will then be combined into a composite index to evaluate whether the interventions were implemented with fidelity.

Teachers were asked whether they attended training for EFAL in 2019. Figure 22 shows that 92% and 91% of teachers in the on-site coaching and virtual coaching interventions responded that they attended training for EFAL, compared to 55% of teachers in the control group. The training received by the teachers in the control group was most likely provided by the province or the district and is considered as the default situation. To double-check this, teachers were also asked whether they received any support for the teaching of EFAL from various organisations. 12% of the control schools responded that they received support from the National Education Collaboration Trust (NECT). The support by the NECT entails providing lesson plans which are very similar to the EGRS lesson plans and cascade training. Fortunately, this is a small number of schools and as a robustness check, we will run the main results on a sample that excludes the schools which received the NECT support.

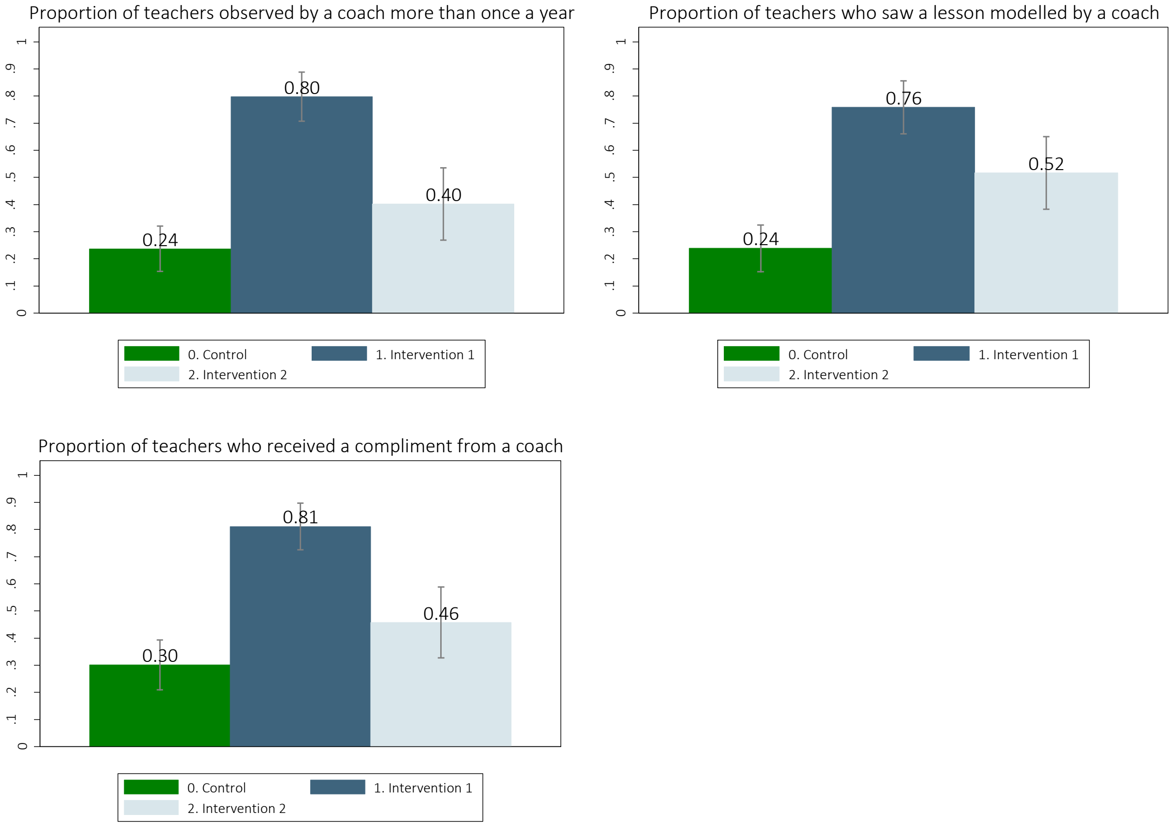
Figure 22: Implementation fidelity – teachers reported that they received training in EFAL



The second check for implementation fidelity is whether teachers received support from the coaches. Support from coaches between the on-site and virtual coach is expected to look different. Being observed by a coach will mean that the on-site coach is in the classroom observing the teacher, but for the virtual coach, it would mean that the teachers submitted videos and photos to the small competitions that were run once every two weeks. Similarly, having a lesson modelled by the on-site coach would mean that the coach modelled a lesson in the classroom, but for the virtual coach, it would mean the teacher watched one of the videos that were developed and sent by the virtual coach. We would need to determine, however, whether the teachers in the virtual coaching group interpreted the virtual support as stated above.

Figure 23 shows that teachers who were supported by the on-site coach were more likely than both the control group teachers and the virtual coaching teachers to respond that they had been observed by a coach, that a coach modelled a lesson for them and that they received a compliment from a coach. Teachers supported by the virtual coach were more likely than the control group teachers to have responded positively to these questions, but this is only statistically significant for the question of whether the coach modelled a lesson.

Figure 23: Implementation fidelity – teachers reported having been supported by a coach


Proportion of teachers observed by a coach more than once a year

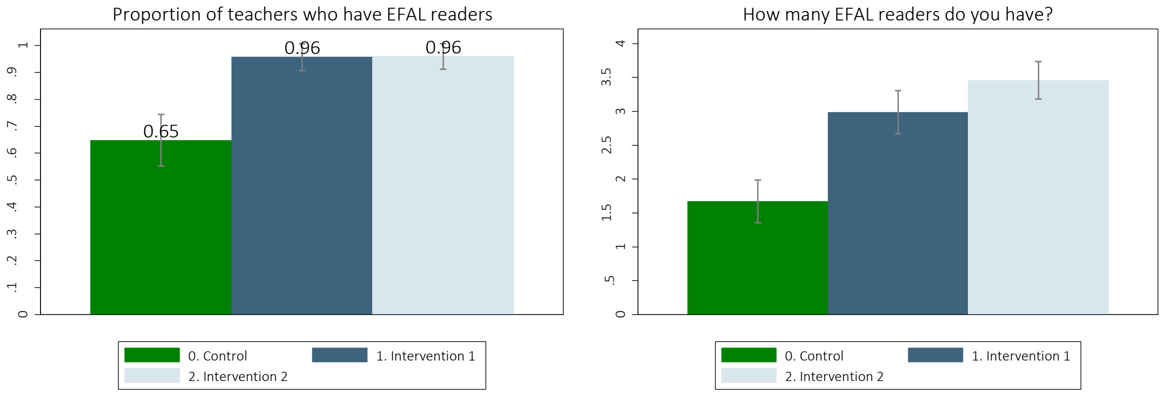
Proportion of teachers who saw a lesson modelled by a coach

Proportion of teachers who received a compliment from a coach

In the third year of implementation, all the Grade 3 learners in the intervention schools were supplied with a graded reader that is an anthology of 25 graded reading titles. Teachers were asked on whether they have EFAL readers in their class and if they had, how many EFAL readers they had in their class.[4] The largest majority of teachers in both intervention groups responded that they had graded readers in their class. They were also more likely to respond that they have a higher number of books in their class relative to the control group teachers.

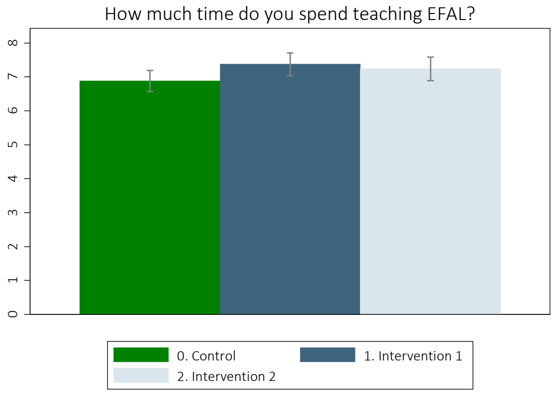
[4] This question gave teachers 5 categories to choose from: 1: 1-9 books; 2: 10-19 books; 3:20-29 books; 4:30+ books; 99: No books.

Figure 24: Implementation fidelity – teachers' access to graded readers



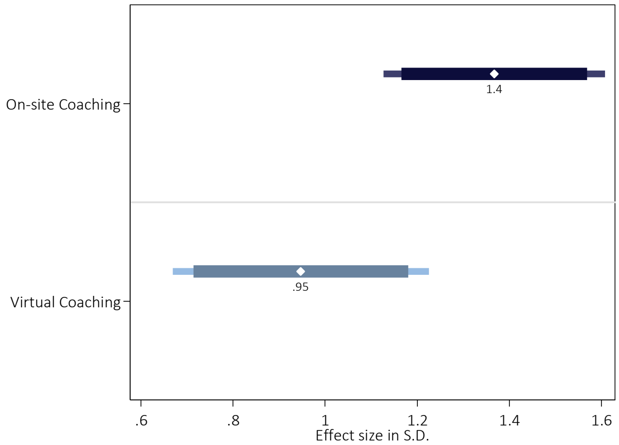
Finally, teachers were also asked what amount of time they spent on average teaching EFAL. This question emanates from the choice the curriculum allows teachers to make between either teaching 3 hours or 4 hours of EFAL a week (and then 7 hours or 8 hours of HL). The EGRS II lesson plans specified that teachers had to spend 4 hours teaching EFAL. The question gave teachers a couple of options to choose from ranging in 30-minute intervals from 1 hour to 5 hours. There was no significant difference between the different groups of teachers on this question.

Figure 25: Implementation fidelity – time spent teaching EFAL



To evaluate the difference in implementation fidelity between the different groups, an index of the components discussed above was created using multiple correspondence analysis (MCA). The index was created using the seven variables which have been discussed in this section. The regression model is shown in table 7 in the Appendix, but figure 26 shows the coefficients for both the on-site coaching and virtual coaching groups.

Figure 26: Coefficients for the Implementation Fidelity Index

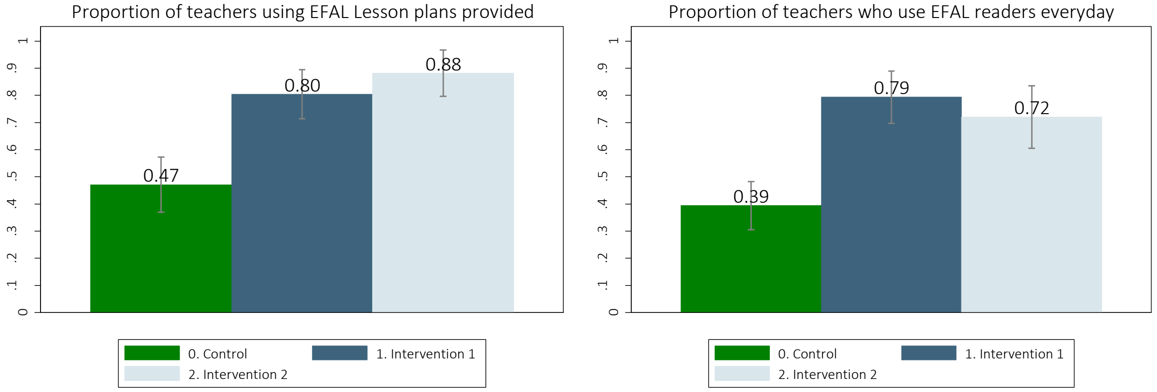


Teacher instructional practice:

The theory of change is that the interventions will encourage teachers to change their instructional practice. The lesson plans are intended to prompt teachers to use a wider range of instructional practices than what they are used to, as well as do these practices more frequently. We will evaluate whether we see evidence of this by considering whether teachers made use of the lesson plans and EFAL graded readers provided. Another way to see whether teachers are familiar with the lesson plans are to see whether they know how often they should be teaching the activities as specified by the lesson plans. Finally, given the broader spectrum of practices that teachers are engaged in, as well as the higher frequency of using these practices, we expect to see evidence of more writing activities in the learners' books. Similar to evaluating the extent of implementation fidelity, a composite index will be created to evaluate the overall extent of teacher instructional practice change between the control and intervention groups.

The first question considers whether teachers used the resources that were provided. Figure 27 shows that teachers in the interventions schools were more likely to report that they used EFAL lessons plans that were provided by either the province or an NGO. We know that the province had developed their own lesson plans and these had been provided to schools prior to the study – this may account for the 47% of control schools who reported that they used lesson plans. Figure 27 further shows that teachers in the intervention schools are also more likely to report that they used the graded readers everyday, which is evidence of teachers using the lesson plans.

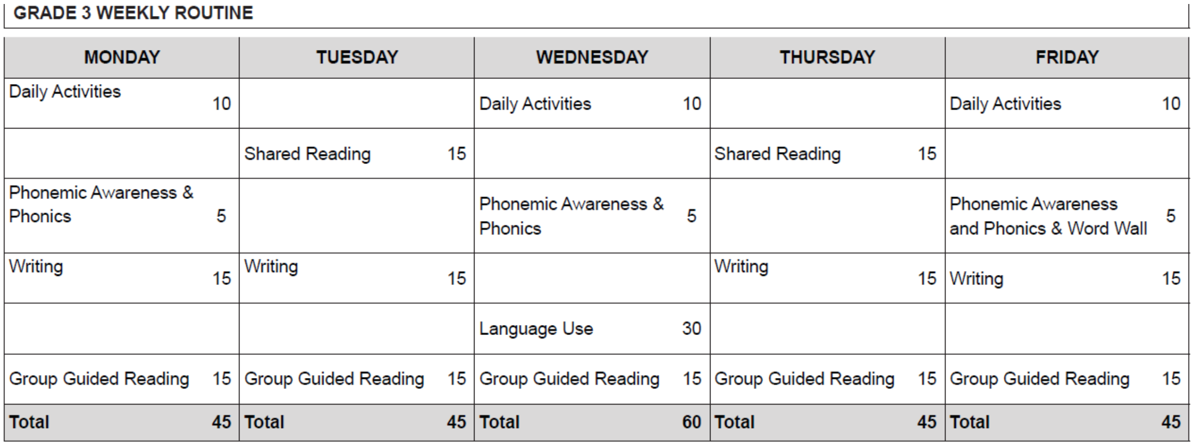
Figure 27: Teacher practice – teachers' use of lesson plans and graded readers



Fieldworkers were also asked to look at the classroom and rate the print richness of the environment on a scale of four, with four being a very print-rich classroom with high-quality materials. Both intervention groups were more likely to score higher than the control schools on the quality of EFAL posters, EFAL flashcards and the availability of storybooks.

Next, we evaluate whether teachers in the intervention groups are more likely to report the correct number of times that the lesson plans specify the teaching of an activity. Table 14 shows the Grade 3 weekly routine and the frequency of each activity that the lesson plans suggest. Teachers in the intervention groups were more likely than the control group teachers to correctly specify the number of times that they should repeat a phonics sound (the core methodologies specify three times), teach phonics (three times a week), teach group-guided reading (every day) and do writing (four times a week).

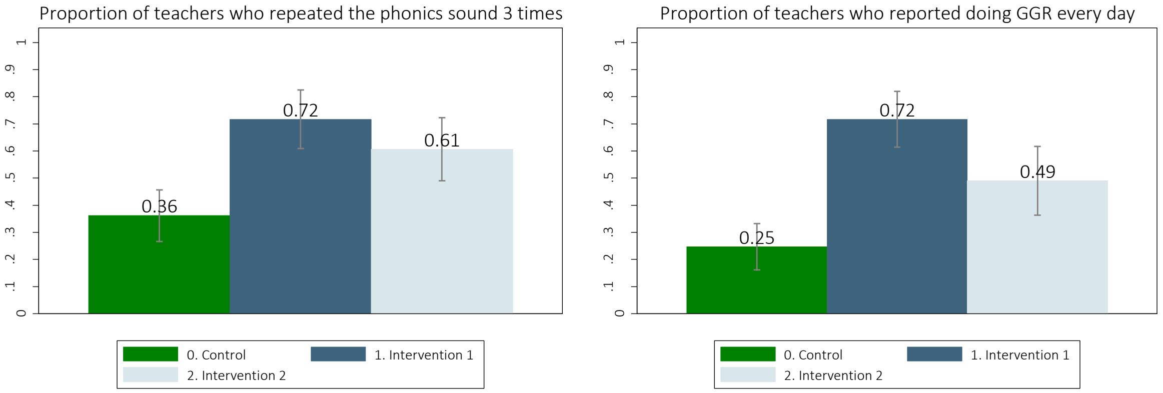
Table 17: Grade 3 weekly routine



GRADE 3 WEEKLY ROUTINE

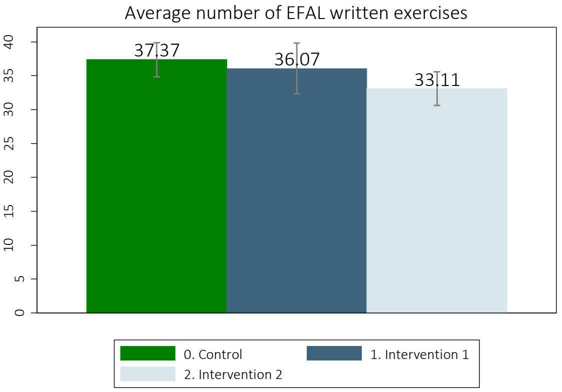
MONDAY		TUESDAY		WEDNESDAY		THURSDAY		FRIDAY	
Daily Activities	10			Daily Activities	10			Daily Activities	10
		Shared Reading	15			Shared Reading	15		
Phonemic Awareness & Phonics	5			Phonemic Awareness & Phonics	5			Phonemic Awareness and Phonics & Word Wall	5
Writing	15	Writing	15			Writing	15	Writing	15
				Language Use	30				
Group Guided Reading	15	Group Guided Reading	15	Group Guided Reading	15	Group Guided Reading	15	Group Guided Reading	15
Total	**45**	**Total**	**45**	**Total**	**60**	**Total**	**45**	**Total**	**45**

Figure 28: Teacher practice – doing the correct frequency of teaching activities



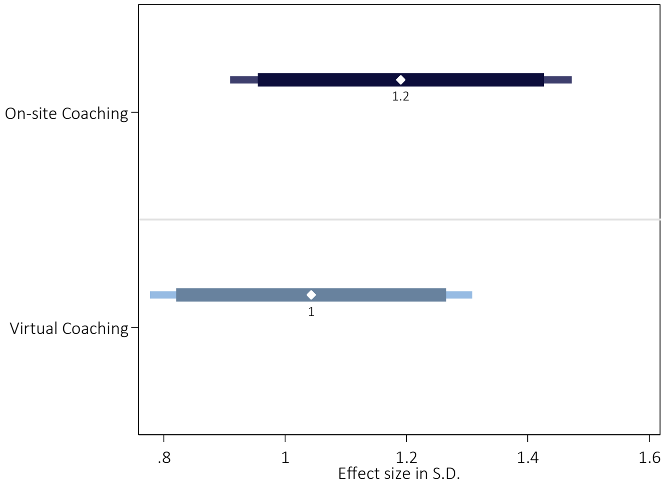
Finally, as mentioned above, we expect to see learners doing more written activities if the implementation of the lesson plans leads to teachers doing writing activities more frequently. Figure 29 below, however, shows that learners in the intervention groups on average did slightly fewer writing activities. This is rather curious but may suggest that the default practice in the control schools is for teachers to do the more familiar instructional practices such as shared reading and writing, at the cost of more difficult practices such as phonemic awareness and group guided reading. Control school are also more reliant on the DBE workbook as their primary resource for teaching, and the workbooks predominantly provide writing activities. The information from the classroom observation study will interrogate this finding further.

Figure 29: Teacher practice – number of written activities completed.



All the aspects of teacher instructional practice change that we looked at in this section are included in the construction of a teacher instructional practice index. Again the index is constructed using multiple correspondence analysis (MCA). Figure 30 shows the coefficients of a regression run on the index and it is evident that teachers in the virtual coaching group changed their practices by one standard deviation and teachers in the on-site coaching group by 1.2 standard deviations. Table 8 in the appendix shows the full regression table for the index and its underlying components.

Figure 30: Coefficients for the Teacher Practice Index

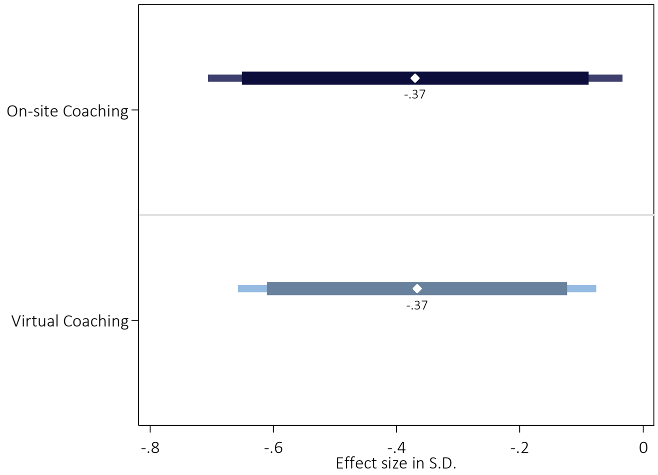


Skill acquisition:

The purpose of the teacher training and coaching was to support teachers with the enactment of various teaching practices. Each term the teacher training and coaching sessions focussed on supporting teachers with specific teaching methodologies. The videos that were given to the virtual coaching teachers also aimed to explain and demonstrate the various teaching methodologies. The expectation is therefore that teachers in the intervention groups will report that they experience less difficulty with the core teaching methodologies relative to the control group teachers. To evaluate whether this was indeed the case, teachers were asked to rate the difficulty that they experience with each teaching methodology on a rating scale ranging from 1 (very easy) to 4 (very hard). Lower average scores therefore indicate that teachers found the methodologies easier.

On average teachers in both intervention groups were more likely to report that the methodologies were easier. The largest difference was reported for the phonics methodology with teachers in both the virtual coaching and on-site coaching groups being more likely to report that the methodology was easier. However, there was no significant difference in the responses of teachers from the different groups on the difficulty experienced with the group-guided reading methodology. Figure 31 shows the coefficients on the index score and table 10 in the Appendix shows the full regression table.

Figure 31: Coefficients for the Skills Acquisition Index



School Support:

The final mechanism through which the interventions may affect instructional practice is through additional support by the school management team members or communities of practice. Although not a primary outcome of the programme, the principal and head of department of the foundation phase were invited to each teacher training session so that they are aware of the methodologies that teachers are implementing. During these training sessions, a separate session was held with the SMTs specifically to encourage and equip them to provide more regular support to the teachers in the intervention. SMT members were invited to the training session from Year 1 of implementation, so provided that they have been SMT members over the past 3 years, they would have received three years of exposure to the programme. Further to this, the on-site coaches also made an effort to check-in with the principal or Head of Department (HOD) every time they visited a school. Similarly, the virtual coach communicated regularly with the SMTs. Finally, both the on-site coaches and the virtual coaches encouraged the teachers to form communities of practice to prepare for the weekly lessons together or to support each other with the more difficult instructional practices.

Despite the effort to get SMT members more involved, we saw from the service provider reports that SMTs in the virtual coaching intervention were less likely to attend the training sessions than the SMTs from the on-site coaching intervention. The virtual coach also reported that the SMTs were not very likely to respond to the communication from her. We would therefore expect to see that SMT members in the on-site coaching intervention provide more support to teachers than SMT members in the virtual coaching and control schools.

Similar to the questions that were asked to teachers about the support they receive from a coach, teachers were asked whether they have been supported by their principal and HOD in the following three ways: the SMT observed a lesson, modelled a lesson and gave a compliment. Overall, there was no evidence of any changed behaviour by HODs, principals or subject advisors.

Teachers in the on-site coaching observation were slightly more likely to respond that they have been observed by their HOD, but this is only significant at a 90% level (table 11 in the Appendix).

We also asked teachers whether they have cluster meetings or meet as communities of practices about teaching EFAL. Figure 32 shows that a slightly higher proportion of teachers that received on-site coaching answered positively to this question. However, the difference was not statistically significant.

Figure 32: Teachers are in a Community of Practice

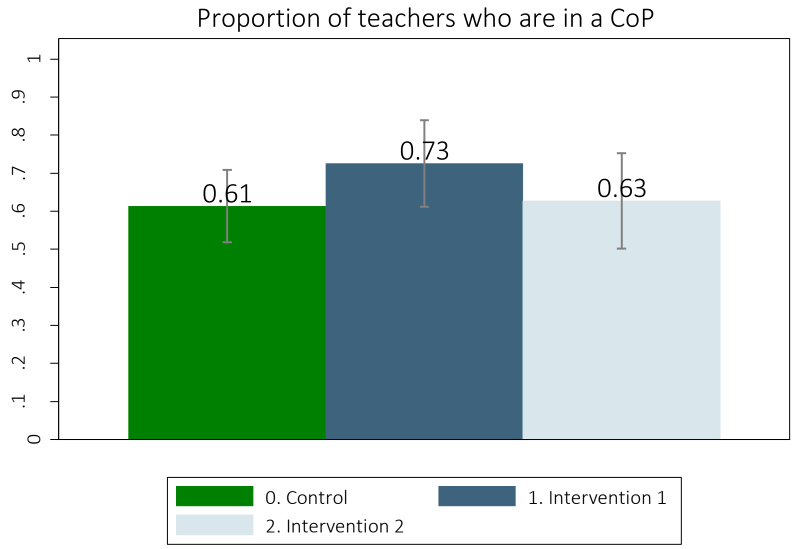


Figure 33: Coefficient of the SMT support received

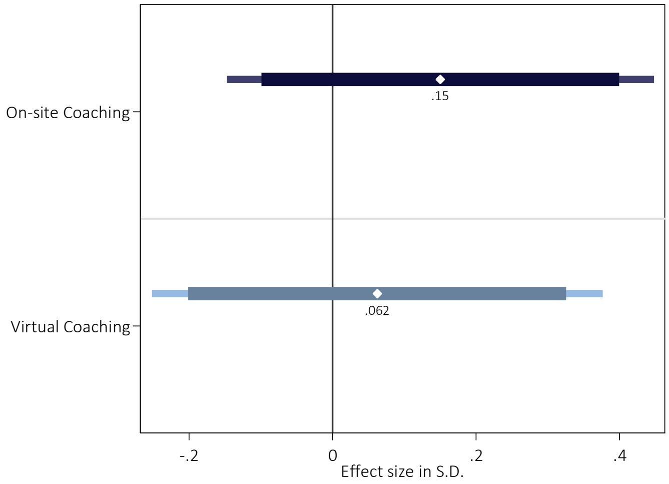
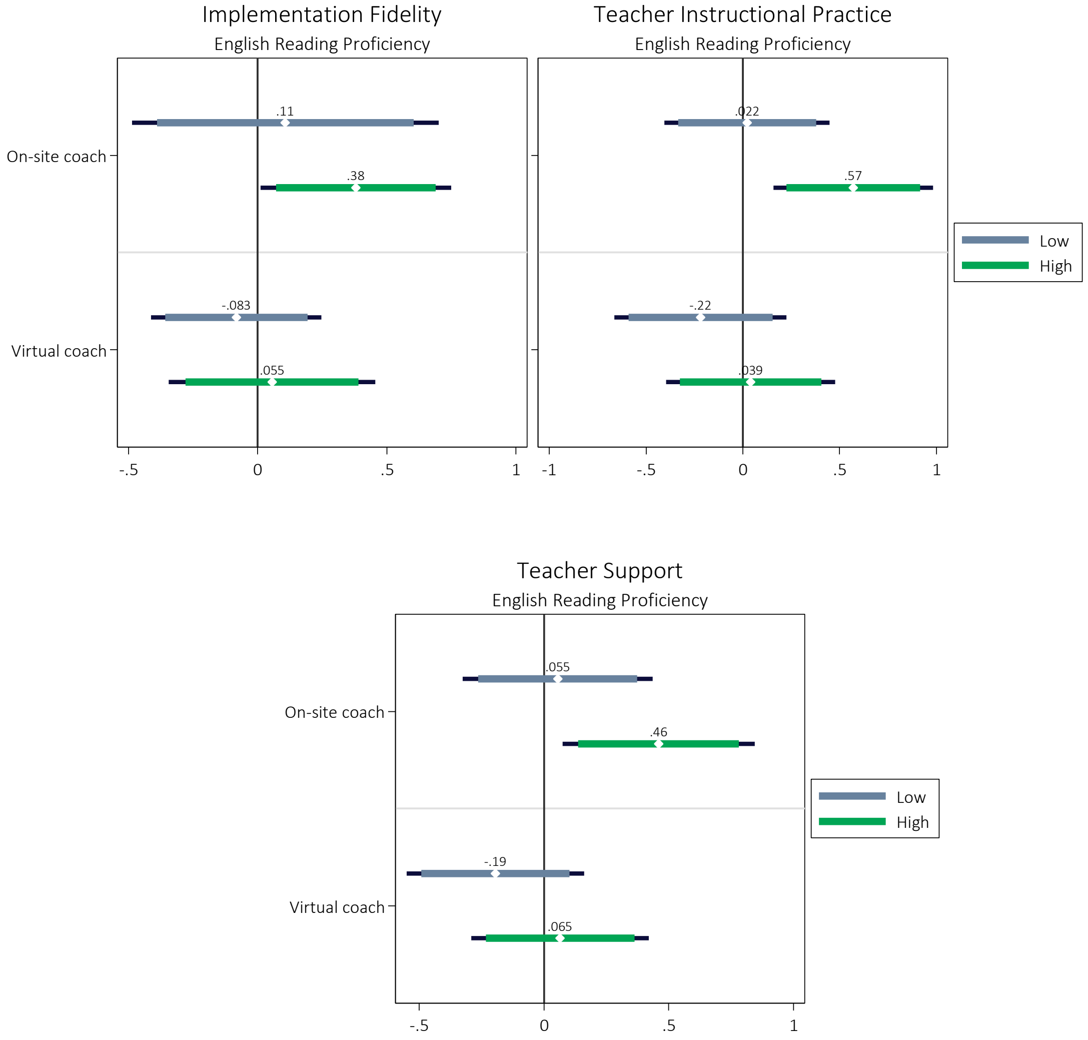


Figure 34 below shows the differences in the coefficients on learner reading proficiency between the teachers who scored either in the top or bottom half of the index distribution. Although the interaction effects are stronger for teachers who scored higher on each index, there is never a significant difference between the two groups. When running the main regressions including interactions with the indices above, no significant interactions were found either.

Figure 34: Interactions with implementation fidelity, teacher instructional practice and in-school support to teachers



5.4. Tablet usage

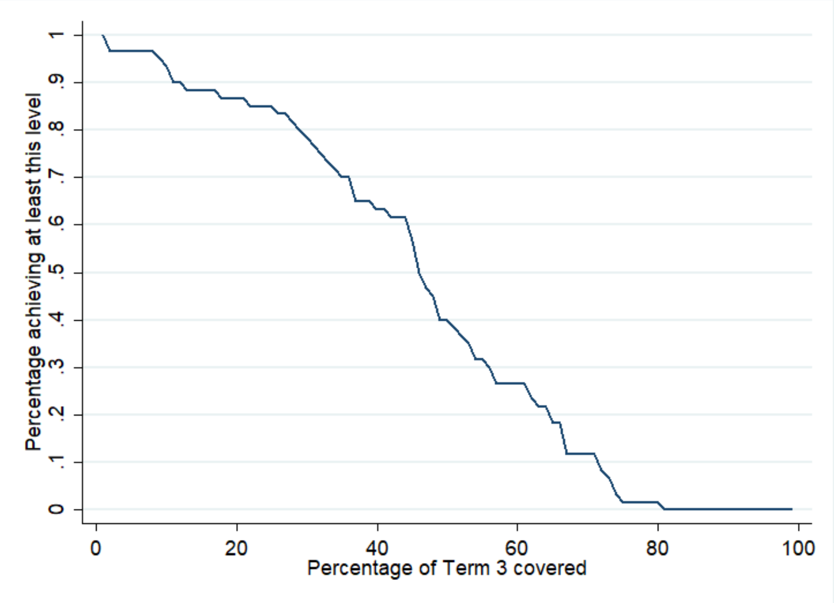
Adherence to the learning programme was difficult to reliably observe. However, for the virtual coaching intervention, we have access to rich tablet usage data, which has records of every occasion teachers accessed any particular slide or content on the tablet. Due to some challenges in extracting this data, the most complete dataset exists for term 3 of 2019.[5]

Figure 35 shows the percentage of term 3 slides that were accessed by teachers any time between July and September. This might be considered a crude measure for potential curriculum coverage, or alternatively, a proxy for intervention implementation fidelity. Evidently, there was quite a range

[5] For a description of the problems with matching the tablet usage data, please see the note in the Addendum.

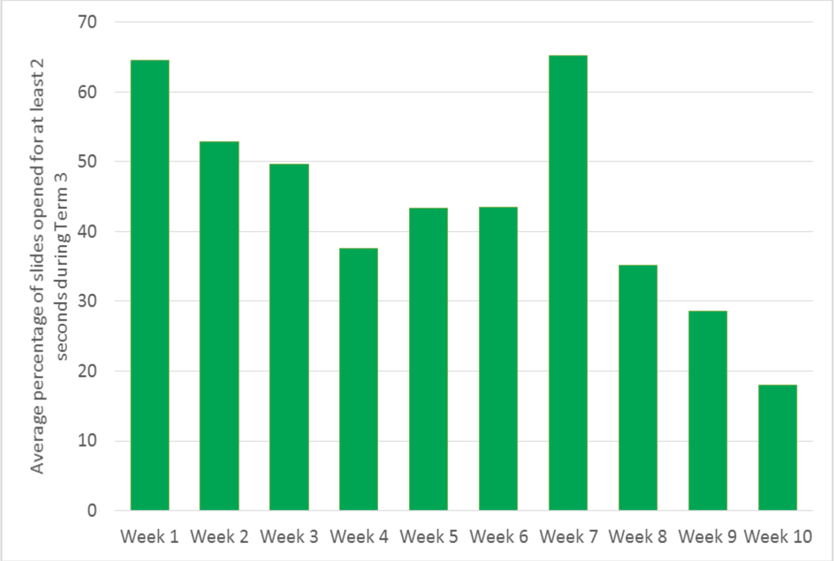
of slide usage across teachers. About 65 percent of teachers reached at least 40 percent slide coverage, while only 27 percent of teachers covered more than 60 percent of the term's slides.

Figure 35: Percentage of term 3 slides covered



A breakdown of slide coverage by each week of term 3 was even more revealing (Figure 36). Interestingly, week 7 was particularly well covered, and this is almost certainly because this is the week in which assessments must take place. Teachers are expected to upload assessment results onto SA-SAMS, a government-wide school management system into which teachers have to upload various data. It is also interesting that, aside from week 7, there seems to be a pattern of better coverage earlier in the term – weeks 1, 2 and 3, with a steady decline in coverage until weeks 9 and 10 which had the lowest levels of coverage.

Figure 36: Average percentage of Term 3 slides covered, by week



The fact that teachers were able to access slides in week 7 (when it might have been perceived to really matter) and the pattern of steadily declining coverage through the term, would suggest that the technology itself was not the main barrier to programme implementation, but rather other factors such as the motivation of teachers or their ability to keep pace with the curriculum (and this not necessarily due to their own fault but quite possibly due to other challenges such as disruptions to schooling beyond their control). This is an important point, since if it can be accepted, it would imply that the reason for the virtual coaching intervention being less effective than the on-site coaching is less likely to be the format of the lesson plans, and more likely to be linked to the coaching model. It is possible, for instance, that on-site coaches were better able to monitor curriculum coverage or to help teachers mitigate the consequences of disruptions and delays in the learning programme.

One of the concerns with implementing a technology-based intervention is on teachers' levels of comfort in working with technology, and particularly so for the older teachers. From interviews with teachers in the case study by Alsofrom (forthcoming) it emerged that despite little experience in working with tablets or computers, teachers overcame these fears relatively easily. This is further evident when comparing tablet usage from teachers in different age categories. The teachers who were older than 55 years on average covered as much of the term 3 curriculum using the tablet as teachers who were younger than 45 years old.

Figure 37: Tablet usage by teacher age

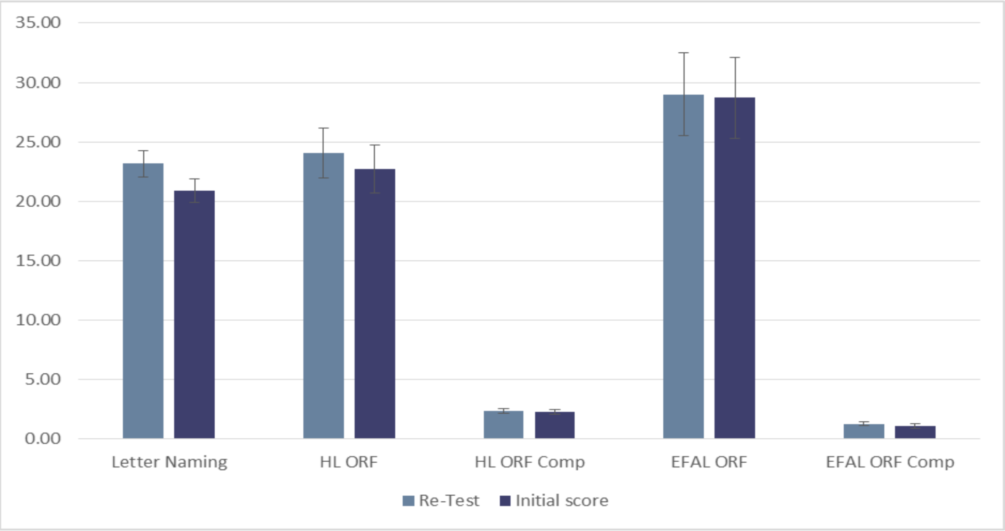
	# Teachers in each category	Average hours	Total slides covered	Average coverage of term 3
Younger than 45	9	17.6	567.5	50.1
45-54 Years Old	26	13.6	550.1	43.7
Older than 55	10	22.9	559.9	53.4

5.5. Robustness checks

5.5.1. Retest data

The purpose of the retest was to determine the extent of inter-rater reliability. Although we do control for fieldworker effects in our models, it is useful to get a sense of the variance in scores between two fieldworkers. Figure 38 shows the average scores on the five sub-tasks for the 315 learners who participated in both the main data collection and the re-test. It is encouraging to see that there is no significant difference between the average scores between the re-test data and the initial scores. It is only for the rapid letter naming task that we see more variance between the two scores, but we do not use letter naming in the evaluation of the success of the interventions. Table 12 in the appendix also shows that the correlation between the test and re-test data and indicates that we can be confident in the test scores in the main data collection. The correlations for the oral reading fluency tasks is at 0.92 for both EFAL and HL and slightly lower for the comprehension questions.

Figure 38: Inter-rater reliability –differences between test and retest data



5.5.2. Vocabulary study results

At the end of Year 1 and Year 2, we saw that the EFAL interventions have had a larger impact on the language proficiency outcomes than on the decoding outcomes. One concern with this outcome was that the vocabulary tasks that we administered were very brief and the vocabulary assessed were focussed on one theme specifically (farm animals). We therefore decided to include an extended vocabulary assessment to determine whether learners' overall English vocabulary increased, and not only the vocabulary that they would have come across in the workbooks. The test has been carefully designed using the frequency distribution of the words in English.

Similar to the outcomes in the short vocabulary test, we see that there was a significant increase in the vocabulary of the learners in the on-site coaching intervention, but not in the virtual coaching intervention. In terms of raw scores, the learners in the on-site coaching intervention could correctly identify 4.5 more vocabulary words and learners in virtual coaching intervention 2.6 words more. The extended vocabulary sample is much smaller than the main sample, which means that the precision of the point estimates is also lower.

Table 18: Coefficients on the English extended vocabulary task

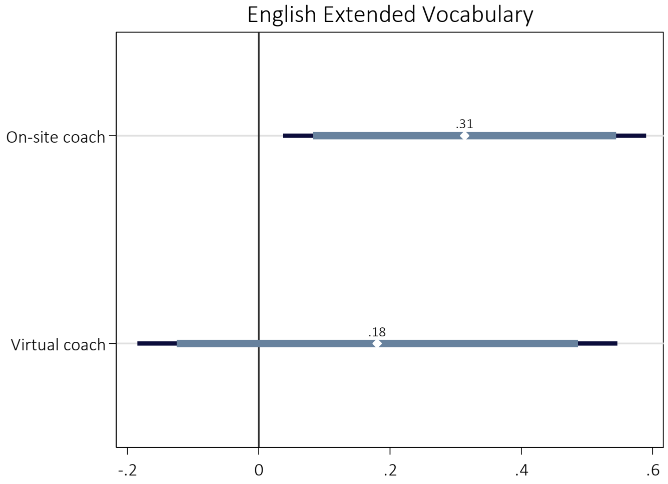
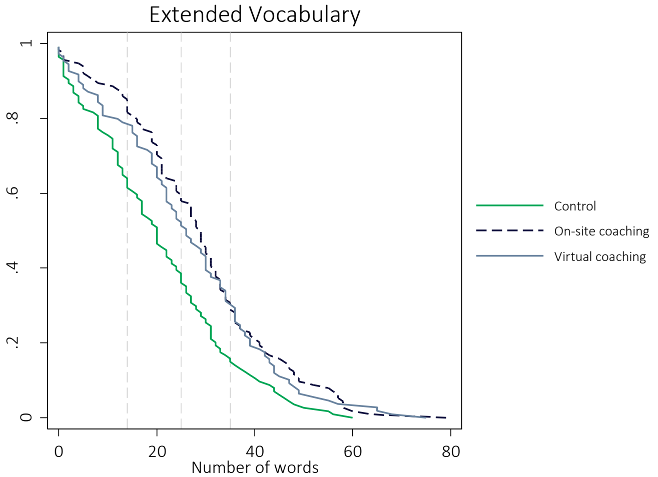


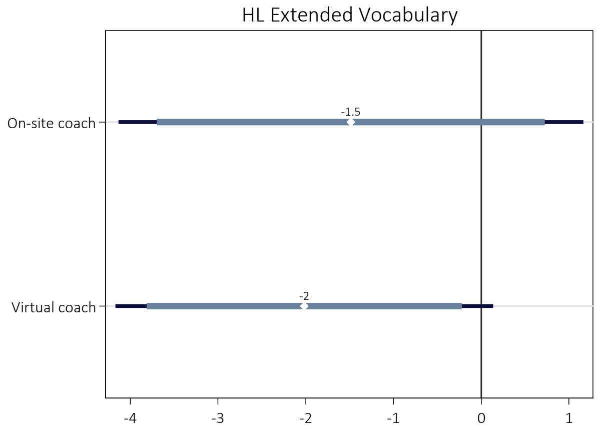
Figure 39 shows the difference between the groups across the performance distribution. At the lower end of the distribution, learners in the virtual coaching intervention performed marginally better than the control learners, but in the top half of the distribution, we again see very little difference in the performance of learners in the on-site coaching and virtual coaching interventions.

Figure 39: Extended vocabulary performance distribution



The HL vocabulary assessment further supports the results that we have been finding for the learning outcomes in HL. On average, the learners in both the on-site coaching and virtual coaching interventions scored lower on the HL vocabulary assessment, although this was only statistically significant at the 90% level for the virtual coaching intervention.

Figure 40: Coefficients on HL Extended Vocabulary



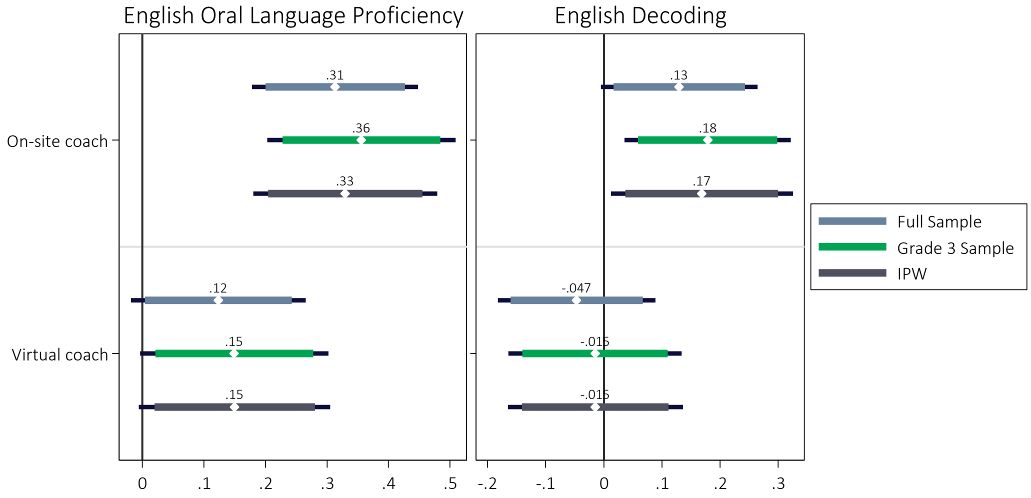
5.5.3. Correcting for attrition and learner repetition

The analysis in our report has been on the sample of learners who received the maximum dosage of the interventions – that is, the learners who were in the classes of the teachers who received the interventions each year (Grade 1 in 2017, Grade 2 in 2018 and Grade 3 in 2019). Excluding learners who repeated a grade will only be valid if we are confident that grade repetition was random. However, as noted in section 3.3, attrition in the sample does not seem to be systematically correlated to treatment status, but we do see that learners in the virtual coaching intervention were more likely to have repeated either Grade 1 or Grade 2.

Figure 41 shows the implication of three different sample specifications on the regression coefficients of the two main indices. The first specification is the full sample, which includes the learners who were found to be in either Grade 1 (very small percentage) or Grade 2 in 2019.[6] The second specification is what we have used in the report and includes only the learners who were in Grade 3 in 2019. The third specification includes inverse probability weights which reweight the data in the Grade 3 sample to correct for the probability of learners that may have repeated or attrited from the sample.

[6] These learners were assessed on the same Year 3 assessment.

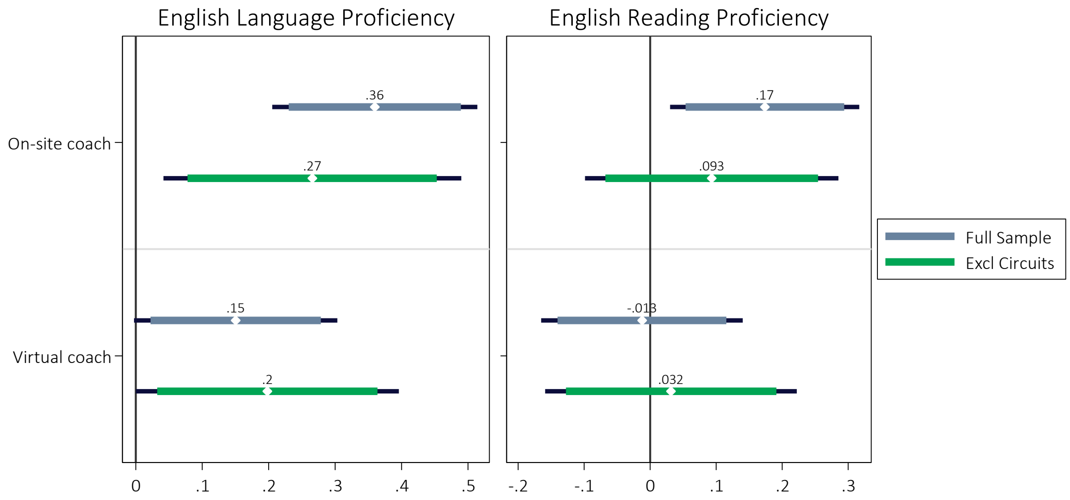
Figure 41: Comparing the effects of attrition and repetition on the estimation models



5.5.4. Industrial action

The final sensitivity check is to see the effect of the teacher strike in the one education circuit. In the figure below, we have excluded the affected education circuits as they were identified by the coaches. The region affected was rather large and the sample that excludes these circuits leads to a reduction of 720 learners (almost a third of our Year 3 sample). The coefficients in both oral language proficiency and reading proficiency change, but the reduction in the sample also leads to lower precision. It is therefore hard to determine the exact effect of the strike. However, given the randomisation of the interventions, we know that control schools and intervention schools were affected equally and we do not have any reason to expect that the strike affected the intervention schools more than the control schools.

Figure 42: Coefficients of regressions excluding circuits affected by the strike



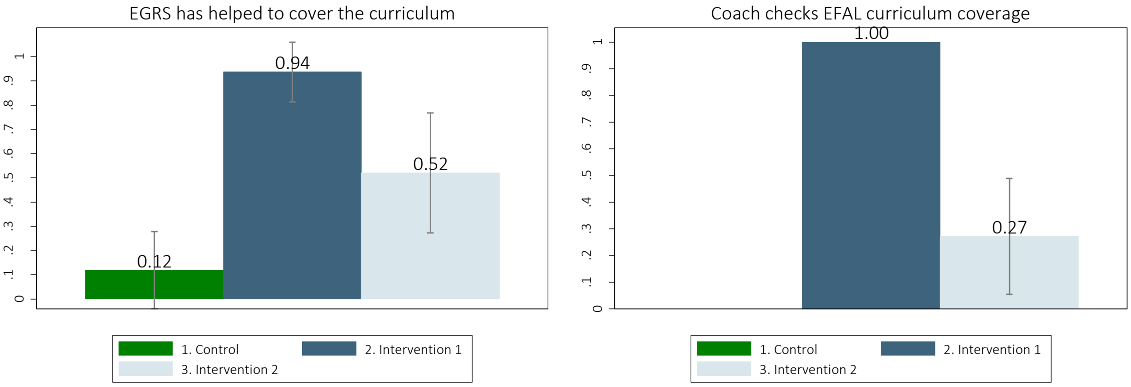
5.6. Classroom Observation Findings

A comprehensive report is available on the findings of the Classroom Observation study. In this section, we will just highlight the main findings as they relate to the mechanisms which are driving change.

5.6.1. Implementation fidelity

The classroom observation study findings support what was found in the main data collection in terms of implementation fidelity. Firstly, the findings in the COS in terms of teacher training were the same. Secondly, teachers were asked what has helped them most to cover the curriculum, and in Figure 43a we see teachers in the on-site coaching intervention more likely to respond that the EGRS has helped. Teachers in the virtual coaching intervention were more likely to respond that EGRS has helped than the control teachers but were almost half as likely to respond "EGRS" than the on-site coaching teachers. To explore this response further, teachers were asked who checked whether they are covering the EFAL curriculum. Again this is quite difficult for the virtual coach to do (since she does not get in the classroom), but various attempts were made to see whether the virtual coach can check curriculum coverage, rather than be dependent on the teacher's responses. Figure 43b shows that the teachers in the on-site coaching intervention unanimously responded that the coach checks their EFAL curriculum, but in the virtual coaching intervention this has not the case. This suggests that the accountability that comes with a coach being in the classroom is more effective in supporting teachers to cover their curriculum.

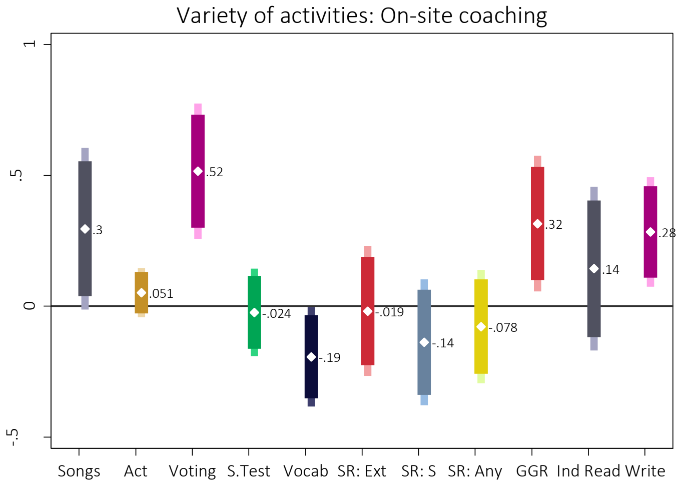
Figure 43: Implementation Fidelity - intervention support



5.6.2. Teacher instructional practice

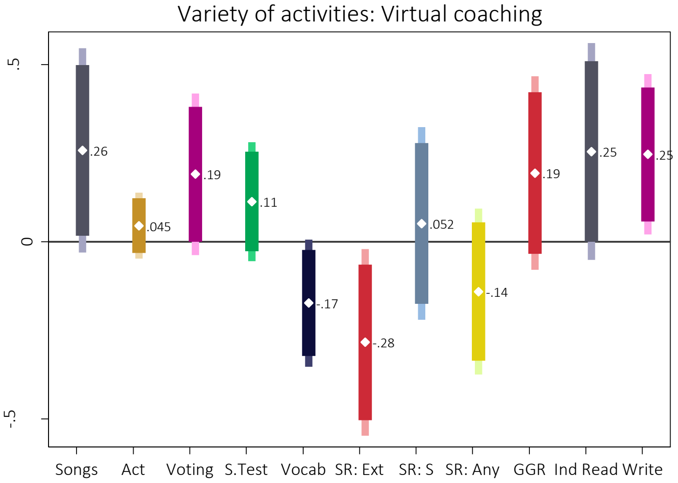
CAPS calls for a variety of instructional practices that are meant to build off of one another progressing through the various levels of phonemic awareness that lead to fluency and the comprehension (CAPS EFAL, 2011, p.15). The lesson plans, being completely aligned to the curriculum, are intended to prompt teachers to use all the methodologies which will result in teachers making use of a wider range of instructional practices than what they are used to. Figures 44 and 45 show the coefficients of regressions run on the probability of observing a range of activities during the lessons. Figure 44 shows that the singing of songs and rhymes (the first grey bar) was observed in 30% more lessons than in the control group. Similarly, the playing of the 'Voting game' (a game that was specified in the lesson plans to assist with language use) were observed in 52% more lessons than in the control group. Figure 44 therefore shows that activities such as acting out stories (gold bar), spelling tests (green bar), shared reading of extended texts (red bar), shared reading of shorter texts (light grey bar) and individual or paired reading (last grey bar) were equally likely to have been observed in the control lessons and the on-site coaching lessons. The activities which were more likely to have been observed in the on-site coaching lessons include the singing of songs and rhymes, the voting game, group-guided reading and writing.

Figure 44: The variety of activities observed during the on-site coaching teachers' lessons



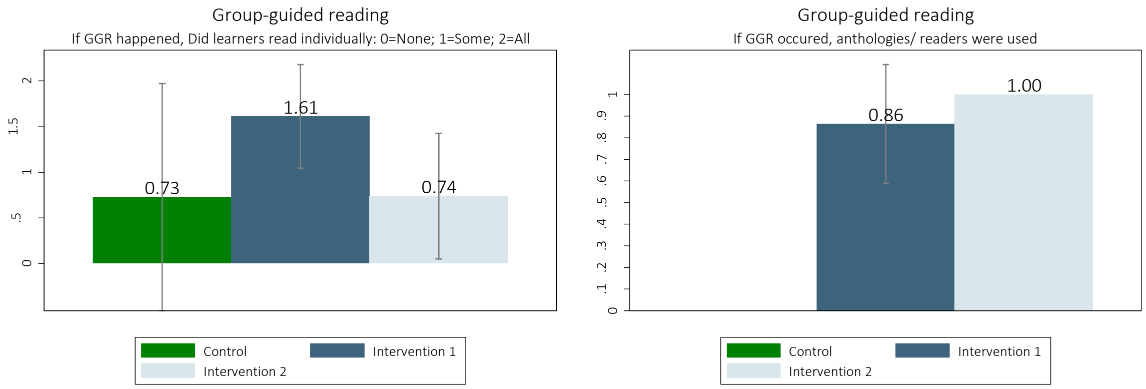
In both the on-site coaching and virtual coaching groups we saw that vocabulary development was less likely to have been observed during the lessons than in the control group. In the virtual coaching group, shared reading of extended texts was also less likely to have been observed. Similar to the on-site coaching group, we also observed that the singing of songs and rhymes, the voting game and writing were more likely to have been observed in the virtual coaching lessons. During the virtual coaching, we were not more likely to have observed Group-guided reading, but we were more likely to observe independent reading happening.

Figure 45: The variety of activities observed during the virtual coaching teachers' lessons



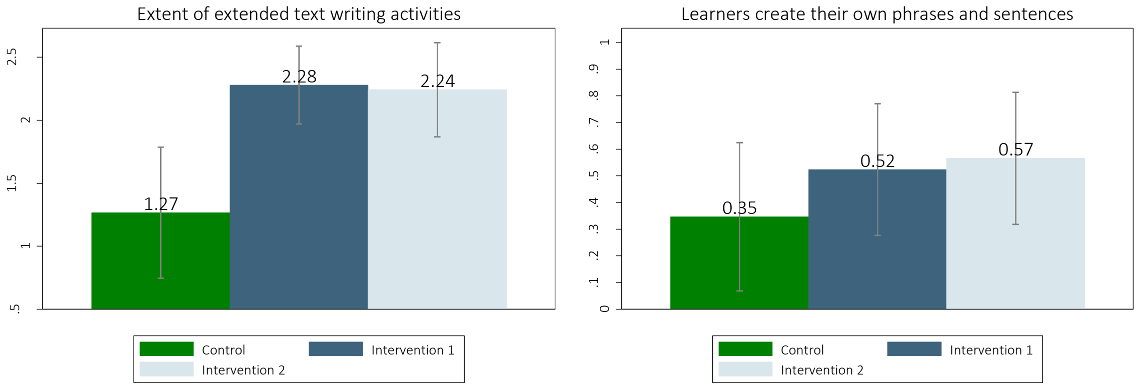
Next, we turn to the rigour in which the activities were implemented. When group-guided reading was observed, the CAPS compliance among the intervention teachers were better on measures such as learners reading from the correct resource (a graded story reader or anthology of stories) and learners having the opportunity to read individually to the teacher (in the on-site coaching lessons specifically). Incidents of questioning and encouraging learners to explain back using their own words were low in both the control and intervention lessons.

Figure 46: Instructional Practices: Group guided reading



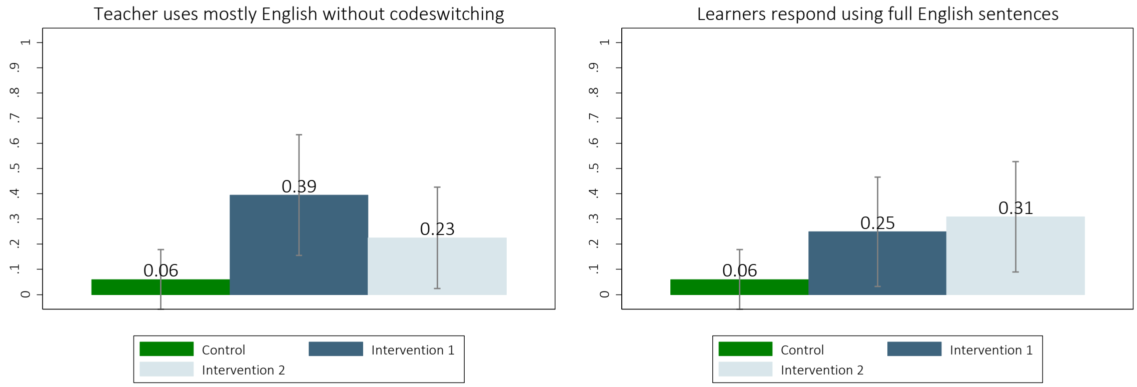
Writing activities were observed more frequently during the lessons of the teachers in the intervention schools. It is also encouraging to see that teachers in the intervention schools were more frequently observed giving their learners writing activities that required them to create their own phrases and sentences (relative to copying sentences or writing only words). Looking at the learners' exercise books, extended writing was also more frequently seen in the intervention schools than the control schools.

Figure 47: Instructional practice: Writing activities



English is a second language for both the learners and the teachers, which means that both parties are not comfortable with using English regularly. The coaches made an effort to encourage teachers to use English more regularly in the classroom and to code-switch only when absolutely necessary. The coaches also encouraged the teachers to get their learners to respond in English regularly. Figure 48 shows that more teachers in the on-site coaching intervention were observed to only use English during the lesson, without codeswitching. Albeit still low, the learners in the intervention schools were also observed responding in English more regularly than the learners in the control schools.

Figure 48: Instructional practices: Use of English during the lesson

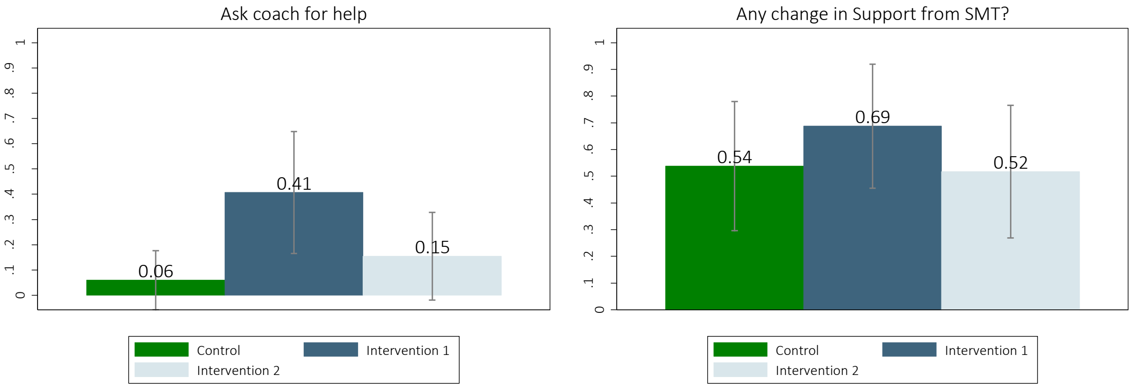


5.6.3. School support

The role of a coach is two-fold: they play a monitoring role by check teachers' compliance with the implementation of the curriculum, but more importantly, they play a support role. For coaches to successfully support and encourage teachers, a trust relationship must develop between the coach and the teacher (Alsofrom, 2019). Measuring whether this trust relationship has developed is difficult, and we attempted to capture this by asking teachers who they most likely turn to for help with an activity or lesson that they find difficult to teach. Teachers were provided with a list of responses from which they had to choose one. 41% of the on-site coaching teachers responded that they would turn to their coach for help, whereas only 15% of the teachers in the virtual coaching intervention responded the same. These results suggest that although the on-site coach only sees her teachers once every three weeks, the fact that the visit is a face-to-face visit helps with building a stronger trust relationship. The virtual coach has the benefit of being able to communicate with teachers any time and day, but it seems as if the lack of having face-to-face interactions does make relationship-building much more difficult.

To see whether the interventions led to any changes in the support that the teachers receive from their SMTs we asked teachers whether they have observed any changes in the support they receive from their HODs and principals. Teachers in the on-site intervention group were more likely to have responded positively to this question. This result supports the findings in the main data collection that teachers were slightly more likely to respond that they have been observed by their HOD.

Figure 49: In-school support



5.7. Case studies findings

Two different case studies were conducted during the third year of implementation. The first case study focussed on the mechanisms of virtual coaching, whereas the second case study focussed on the challenges with implementing an English intervention. Kaitlin Alsofrom conducted the first case study and the Centre for Education Practice Research conducted the second case study. At the time of concluding the report, the second case study has not yet concluded and will therefore only be available in a separate report. The section below draws directly from Alsoform's research report.

The goal of the virtual coaching study was to understand how teachers utilise tablet technology, including virtual coaching, to successfully drive their development and change their teaching practices. Because this question asks what the mechanism is for *successful* teachers, it was critical to interview and observe relatively successful teachers within the intervention. Through a case-study approach utilising classroom observations and in-depth semi-structured interviews, this study explored the critical questions: What are the key support mechanisms through which teachers enact new methodologies? How do teachers learn from the educative materials on the app?

This research study highlighted several very important insights into the mechanisms of change in the virtual coaching intervention. The first was that that comfort with technology was not the key factor in determining which teachers will successfully implement a tech-based intervention. Even teachers who have never used forms of technology such as tablets and computers, were able to embrace a tech-based intervention. Teachers explained that the one reason for this was that the tablet technology felt similar to phone technology that they are comfortable with. The second reason given was that it contained a comprehensive set of pedagogically relevant materials that teachers found very helpful. Given this general ease with technology, teachers were successful in this intervention for reasons beyond just being comfortable with the technology.

Many of the teachers interviewed identified the video technology as key to successfully changing their practices in the classroom. The videos served as easy-to-follow demonstrations of teaching practice. Secondly, the videos seem to help teachers feel supported and reassured because they can use the videos proactively (to remember what they must teach in a lesson) and retroactively (to try out a new practice and then to watch the videos for reassurance).

The competitions that were run by the virtual coach was another critical aspect to the success of the virtual coaching intervention. Teachers winning a competition served as an opportunity to give recognition and praise. When teachers win, they were recognised in a public forum (WhatsApp groups) and they receive praise not only from the coach but from other teachers and their SMTs. The competition submissions also allowed the virtual coach to break down any misconceptions or to clarify problem areas. In this way, the competitions not only provided the coach eyes into teachers' classrooms but also gave teachers eyes into each other's classrooms. This helped to break down the barriers that create the isolated world of the classroom.

Finally, the case study also provided very important insights into why the virtual coaching intervention did not have the same impact on learning outcomes as the on-site coaching intervention. Firstly, the virtual coach did not make in-person classroom visits. She connected to teachers via WhatsApp and phone calls. For teachers who might not be interested in implementing new practices or engaging with their coach, these modes of communication are relatively easy to ignore. Secondly, since there is no strong system of accountability in the tech-based intervention group, the teacher has to intentionally opt into the mechanisms of support (like the competitions) but can easily avoid them if she desires. Teachers can engage more when they choose to and disengage when something feels difficult or uncomfortable. Virtual coaching may therefore work for teachers who are already more proactive and self-motivated because they have to more actively facilitate their own development. The teachers who are successful in this intervention seem to choose to engage and to a certain extent, drive their own development process. Accountability is therefore only accessed by choice.

Alsofrom's findings suggest that because the tech-based intervention inherently includes distance, the barrier to success may be self-motivation. Teachers still felt supported, but there was not a strong enough accountability mechanism to incentivise less motivated teachers to change their practices. It was the teachers who were self-motivated (or were perhaps in an already functional school environment where accountability is provided through principal or colleagues interactions) where the technological intervention seemed likely to be most impactful. Alsofrom concludes that although virtual coaching has the potential to be a cost-effective way to support an abundance of teachers in rural contexts which are difficult to reach, it seems that in dysfunctional school environments with very low accountability mechanisms, this type of intervention is unlikely to be successful.

6. Why are we seeing the effects on the HL items?

In section 5.3.1 we saw that the learners in the virtual coaching intervention had lower scores on the home language proficiency index score than learners in the control group. Similarly, learners in the on-site coaching intervention also scored lower on the HL oral reading fluency task than learners in the control group. This leaves us with two questions:

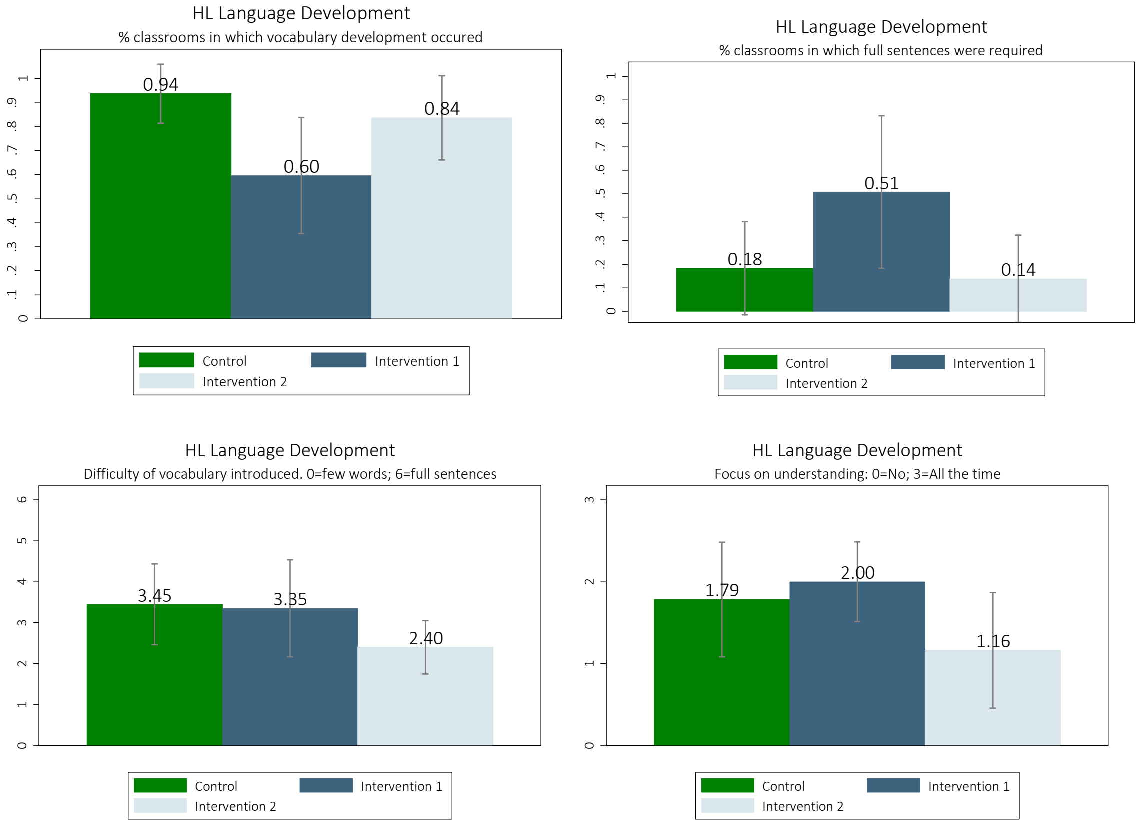
- Did the interventions lead to changes in the instructional practices of teachers when teaching HL, to the detriment of HL teaching?
- Did the interventions lead to an increased focus on EFAL teaching at the cost of HL teaching?

When considering the first question, it is important to note that the interventions did not target HL teaching, but only the teaching of EFAL. No materials were provided for the teaching of HL but in the third year of the intervention, the coaches were encouraged to show teachers where the same teaching methodologies are used in HL and EFAL (for example, Group-Guided Reading, Shared Reading etc). Further, given that the materials and interventions were designed to support the enactment of the EFAL CAPS curriculum, they are completely aligned to the EFAL CAPS curriculum. We did not introduce any methodologies or content that are not specified in the CAPS curriculum. For these reasons, we do not expect that the interventions themselves (i.e. the materials and content) to have had any effects on the teaching of HL.

Looking at the observations of the HL lesson during the classroom observation study, there does not seem to be any clear evidence that teachers' HL teaching practices were affected by the EFAL interventions. No significant differences were observed in teachers' classroom management or time on task, on their teaching of shared reading or in their engagements with learners. Similarly, the intervention group teachers were also not found to be teaching a wider range of activities in HL (Table 13 in the Appendix).

However, we do see that teachers from the interventions schools were less likely to have spent time on HL vocabulary development than teachers in the control schools and this is specifically so in the on-site coaching schools. When teachers did teach vocabulary development, however, we see that the teachers in the on-site coaching intervention are more likely to focus on introducing full sentences with the new vocabulary (relative to the control schools where teachers were more likely to only introduced a few new words). We further see that the teachers in the virtual coaching schools introduced more simplified vocabulary (words and phrases, rather than full sentences) and were less likely to focus on the understanding of the new vocabulary introduced.

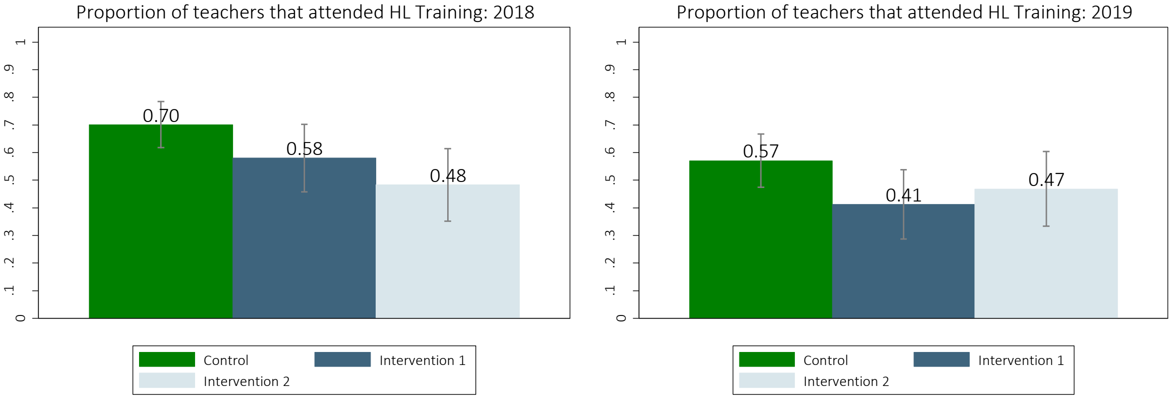
Figure 50: Vocabulary development



Although the teaching of vocabulary development was less likely to have been observed, this was the only evidence of some differences in the instructional practices between intervention and control teachers, and could therefore not have been the mechanism that led to the lower HL outcomes.

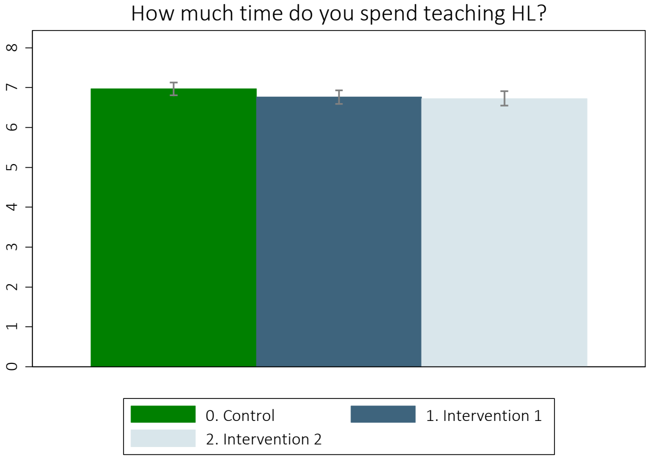
The more likely pathway of impact could be through an increased focus on the teaching of EFAL at the cost of teaching HL. Although we are not able to conclusively say that this crowding-out may have happened, there are a couple of factors which may be supporting this theory. The first is that teachers in both the on-site coaching and virtual coaching intervention were less likely to have attended HL Training in 2018 and 2019. These differences are statistically significant for both interventions in 2018, and only for the on-site coaching intervention in 2019. Various engagements were held with provincial and district officials throughout the intervention period to ensure that they understand that they should avoid implementing EFAL interventions in the selected schools, but that all other training and support to the sampled schools should happen as normal. Nevertheless, it seems as if the intervention schools may have received less support for HL.[7]

Figure 51: Teachers were less likely to receive HL training in 2018 and 2019



The second contributing factor is that teachers in the interventions schools reported having spent less time on average teaching HL. Although we did not see any significant difference between the control and intervention schools in the amount of time that they spend on teaching EFAL (figure 25), we do see a difference on the reported time spent teaching HL (figure 52). The question provided teachers with a selection of response options ranging in 30-minute intervals from 6 hours to 10 hours and on average we see that the

Figure 52: Teachers in intervention groups spent less time teaching HL



[7] The same question was asked in the 2018 teacher questionnaire of the Grade 2 teachers. Although the intervention teachers also reported having attended less training in 2017 and 2018, the differences were not statistically significant.

7. Cost-Effectiveness

The EGRS I study found that instructional coaching as the professional development component of a structured pedagogic programme is more cost-effective than the centralised training model. Building on the findings of EGRS I, the current study investigates the efficiency of alternative models which could be less resource-intensive. The evidence after three years of implementation suggests that the on-site coaching model had a larger impact on learning outcomes than the virtual coaching model.

For cost estimates, the programme budget for the three years of implementation was taken, excluding any costs that were involved in the development and piloting of the programme.[8] These estimates should therefore provide a realistic per-learner cost if these models of delivery are scaled up. Based on these estimates, the per-learner costs of on-site coaching for the three years of intervention is R2,766 and R 2,240 for virtual coaching.[9] This translates to R921 per learner per year for on-site coaching and R747 per learner per year for virtual coaching. In terms of cost per teacher, it is R38,502 per teacher per year for on-site coaching and R31,190 for virtual coaching.

Table 19: Implementation cost

	On-site coaching		Virtual coaching	
	Rand	*US$*[10]	*Rand*	*US$*
Per learner cost for 3 years of implementation	2,766	198	2,240	160
Per learner cost for 1 year of implementation	921	66	714	53
Per teacher cost per year of implementation	38,502	2,750	31,190	2,228

Given the impacts of 0.31[11] on oral language proficiency and 0.13 on reading proficiency for on-site coaching over the three years of implementation, there was a 0.11 standard deviation increase in oral language proficiency for each R1,000 spent and a 0.05 increase in reading proficiency for each R1,000 spent. For virtual coaching, there was no significant impact on reading proficiency, but for oral language proficiency there was a 0.06 increase in oral language proficiency for each R1,000 spent. On-site coaching, therefore, does not only have a larger impact on learning outcomes, but it is also more cost-effective compared to virtual coaching.

[8] Costs such a material revision and the development of new audio sound clips were still included since these activities are done throughout the interventions to respond to the lessons learnt based on the challenges experienced by teachers. These on-going costs will most likely remain in the scale-up of the interventions.

[9] Two costs have been excluded in these figures: An amount equal to 10% of the total costs for overheads and a 15% value-added tax. These costs are the same between the interventions.

[10] All US$ rates are calculated at a Rand:US$ exchange rate of R14 per 1US$.

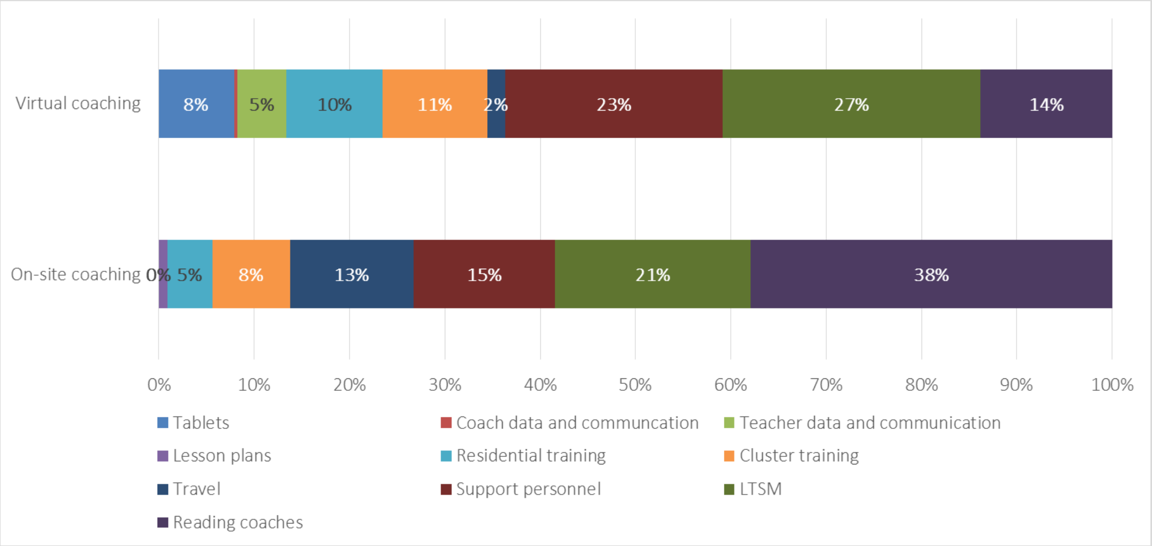
[11] These are the effect sizes on the full sample of learners. We used these, rather than the effect sizes of only the grade 3 sample, since it will give us a more accurate sense of the real cost-effectiveness.

Table 20: Cost-effectiveness of the interventions

	On-site coaching		Virtual Coaching	
	Oral Language	*Reading*	*Oral Language*	*Reading*
Per R1,000 spent	0.11	0.05	0.05	-
Per US$100 spent	0.16	0.07	0.07	-

Figure 53 shows the cost drivers in each intervention. The largest cost item in the on-site coaching model is the salary cost of the four coaches, while the additional night of residential training, the tablets and cellular data for teachers are the main cost items in the virtual coaching model. A critical resource to the quality of both the on-site coaching and the virtual coaching models is the coaches. To support the 86 Grade 3 teachers in the 50 intervention 1 schools, four specialist reading coaches were employed, while one 'virtual' coach was employed in intervention 2 to support a similar number of Grade 3 teachers in 50 intervention 2 schools. The availability of expert reading coaches in each of the country's 11 home languages is therefore an important resource constraint that will need to be taken into account in decision-making regarding the feasibility of taking the coaching model to scale.

Figure 53: Comparison of cost drivers between the two interventions



7.1. Cost of electronic lesson plans

Two innovations were trailed in this study: the electronic lesson plans and the virtual coaching modality. In section 5.4 we saw that the technology was not a problem for accessing the electronic lesson plans, which could make this a viable option, should it be more cost-effective than paper-based lesson plans. Tablets are often thought to be more cost-effective since they can be used for multiple years. However, other yearly costs need to be taken into account, for instance, the costs of hosting the application on a server and the costs of a technical assistant to support teachers who experience technical problems with the tablet. Table 21 below shows a comparison of the costs of the paper-based lesson plans, relative to the electronic lesson plans. These costs assume that the application to host the lesson plans are already developed and needs no additional software development. Table 21 therefore shows that in 2019 it cost us almost R51,000 to provide the paper-based lesson plans to 86 teachers, whereas the electronic lesson plans were almost R550,000. This made the electronic lesson plans 8.8 times more expensive that then paper-based lesson plans.

If we were only to provide teachers with a tablet and the lesson plans in pdf format on the tablet, then the costs would break-even after four years. However, we would probably need to replace teachers' tablets once every four years which means that the tablets and paper-based lesson plans are equally expensive. However, this does not take into account the technical support that teachers will require if the tablet malfunctions or breaks. The other additional expense arises from developing and hosting an application which could make the use of the lesson plans more interactive and supportive. The server and application maintenance cost is about R126,179 per year, and this is a recurring cost. Including these additional costs means that the electronic lesson plans will always be much more expensive than the paper-based lesson plans.

Table 21: Comparing costs between paper-based and electronic lesson plans

Paper-based lesson plans		Electronic lesson plans	
Printing	R 50,926.25	Tablets	R 200,071.00
		Hosting and basic software maintenance	R 126,179.33
		Technical support	R 123,604.40
TOTAL	**R 50,926.25**	**TOTAL**	**R 449,854.73**

Notes: These are costs per year for about 85 teachers

7.2. Cost of virtual coaching

A similar comparison can be made between the yearly costs of supporting teachers with either an on-site coach or a virtual coach. To support the 86 teachers in the on-site coaching intervention, four coaches were employed to ensure that teachers were observed and supported at least once a month. Travel and accommodation costs for these coaches were also quite high, especially for the coaches who supported teachers in the more rural areas. One virtual coach managed to support a similar number of teachers (82) and had minimal travel costs to attend the quarterly cluster training sessions. The additional costs in virtual coaching relate to data and communication cost. The data and communication cost for the coach was about R24,000.00 per year. To enable teachers to send and receive the messages, videos and audio clips from the coach, we also provided each teacher with 1GB of data each month, which worked out to R130,00.00 per year. On-site coaching is therefore 3.4 times more expensive than virtual coaching.

Table 22: Comparing costs between on-site coaching and virtual coaching

	On-site	Virtual
Coach salaries	R 1,351,399.23	R 349,722.52
Coach travel	R 440,871.01	R 27,264.60
Coach communication		R 24,289.80
Teacher data		R 130,510.33
TOTAL	**R 1,792,270.23**	**R 531,787.25**

8. Discussion

This study compared the effectiveness of a structured pedagogy programme that was implemented through two different delivery models. The first was through providing teachers with paper-based lesson plans and support by an on-site coach. The second was through providing teachers with lesson plans on a tablet and support by a virtual coach. The main research question, therefore, considers whether an alternative form of coaching could work. Could virtual coaching combined with lesson plans and other resources on an electronic tablet be an effective alternative to on-site coaching?

On-site coaching has become an accepted model for improving early grade reading, and indeed this study we found the same. Learners in the on-site coaching group saw an effect size of 0.31 standard deviations in oral language proficiency and 0.13 in reading proficiency skills. Virtual coaching, however, had a much weaker effect on learning outcomes, with an impact of less than half of that of on-site coaching (0.12) for oral language proficiency and a negligible effect on reading proficiency.

Since the content in the two programmes were the same, there are two possible reasons for the difference: either the technology (the tablet) was a barrier to teachers implementing the programme or virtual coaching does not have the same efficacy as on-site coaching.

Through the analysis of some of the tablet usage data, we could determine that the largest majority of teachers were using the electronic lesson plans especially at the start of the term and again later in the term during the week when learner assessments took place. This suggests that the model of lesson plan delivery may not have been the reason for weaker learning outcomes, but rather that the coaching model was not as effective as on-site coaching.

To better understand the difference in the mechanics between the two coaching models, we looked at implementation quality, coaching support, instructional practices change, change in teachers' skill acquisition and change in in-school support by SMT members.

In both interventions, we saw high attendance at the teacher training, as well as high use of the lesson plans and the graded readers. Implementation quality in terms of curriculum coverage is harder to observe, but for the intervention 2 teachers, we had rich data in terms of tracking their tablet usage. Using this data we estimate that only 27 percent of teachers accessed more than 60 percent of slides, indicating relatively low curriculum coverage. Coverage was better earlier in the term than towards the end of the term, suggesting that many teachers struggled to keep up with the learning programme either due to a lack of motivation or their ability to keep pace with the curriculum.

Teachers in the on-site coaching intervention also showed stronger efficacy of implementation as seen by their higher likelihood to know the correct frequency of teaching activities, being more often observed teaching the more difficult methodology of group-guided reading and speaking English more often during the lessons. The classroom observations also saw on-site coaching teachers teaching a wider spectrum of the core methodologies and teaching them more often. This suggests that the on-site coach may be able to provide stronger support to teachers in changing their instructional practices.

When teachers were asked about the support that they have received from their coaches, it emerged that teachers in the on-site coaching intervention were more aware of the components of the programme than those in the virtual coaching intervention, and were more likely to have been observed teaching or to have seen a coach modelling teaching practices. The interviews in the case study revealed that since the virtual coach had the challenge of not being in the teachers' classrooms, engagements between teachers and their coach in the virtual coaching programme were more dependent on teachers choosing to make use of the support from the coach (Alsofrom, forthcoming). This suggests that the lack of a clear accountability mechanism in the virtual coaching group could mean that teachers who may be less motivated or may be teaching in environments with less accountability, could easily decide not to implement the programme or change their practices.

From the secondary analysis, it emerged that there may have been a negative impact on HL outcomes. This may either have resulted from teachers having changed their instructional practices to be more detrimental to HL teaching or because of the crowding out of HL teaching time as a result of the additional focus on EFAL. We saw no strong evidence to suggest that the teachers may have changed their instructional practices in HL teaching. We saw, however, that the teachers in the interventions were less likely to have attended HL training and may have spent less time teaching the HL lessons. This crowding-out may have led to the lower learning outcomes in HL teaching.

This study has confirmed the potential for structured pedagogy programs to significantly improve learning outcomes when supported by on-site coaching. However, the main finding of this paper is sobering: a virtual coaching alternative, which was somewhat less expensive and considerably less reliant of human resources, did not have the same effect. The research agenda to design innovative programs that allow meaningful support to teachers at a large scale must continue. But for now the evidence indicates that interventions with a strong theory of change, which may be relatively costly, are needed to start reducing the substantial learning gaps that exist in countries like South

Africa. This is not a convenient finding in contexts that have tight fiscal constraints or where re-prioritisation of public finances is difficult. However, in most education systems the wage bill accounts for upwards of 80 percent of education spending, and in these settings some degree of re-prioritization towards coaching is likely to improve the effectiveness of teachers, and in turn make overall education spending more cost-effective.

Appendix

Table 23: Item descriptive statistics - Full sample

	N	Mean	s.d.	p10	p25	p50	p75	p90	Min.	Max.	% zero score
Letter Naming TC	2684	20.7	9	8	15	21	27	33	0	36	1%
Letter Naming CPS	2684	1	0.5	0.4	0.8	1	1.4	1.6	0	3	1%
Letter Recognition	2684	44	22.4	13	28	45	60	72	0	110	2%
HL ORF at 60 seconds	2684	22	17.8	0	1	23	36	47	0	58	25%
HL ORF Comp	2684	2.3	1.9	0	0	3	4	5	0	5	32%
EFAL Word Recog	2684	23.8	21.9	0	1	22	40	55	0	99	21%
EFAL ORF at 60 secs	2684	28.3	30.4	0	0	19	49	72	0	126	26%
EFAL ORF Comp	2684	1.1	1.5	0	0	0	2	4	0	5	54%
EFAL Produc. Vocab	2684	3.3	1.8	1	2	3	5	6	0	6	7%
English Comp	2684	1	1.1	0	0	1	1	3	0	4	39%
HL Written Comp	2669	2.3	1.9	0	0	2	4	5	0	6	28%
EFAL Written Comp	2669	1.4	1.2	0	0	1	2	3	0	4	31%
Mathematics	2669	1.4	1.2	0	0	1	2	3	0	4	31%
PCA: EFAL Lang Prof.	2684	0.2	1.3	-1.5	-0.7	0.1	0.8	2	-1.8	3.4	
PCA: EFAL Read Prof.	2632	0.1	1.9	-2	-1.6	-0.4	1.4	2.9	-2	5.8	
PCA: HL Read Prof.	2632	-0.1	1.8	-2.6	-2	0.2	1.5	2.1	-3	4.2	

Table 24: Tasks means in Wave 4, by intervention group - Full sample

	Control	On-site	Virtual	C vs On-site	C vs Virtual
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
Gr3 Letter recognition	42.947	48.135	41.657	-5.188**	1.290
	[1.283]	[1.664]	[1.568]		
Gr3 HL ORF at 60 seconds	23.091	21.546	20.684	1.545	2.407
	[0.853]	[0.776]	[1.232]		
Gr3 HL ORF Comprehension	2.401	2.313	2.124	0.088	0.276*
	[0.088]	[0.103]	[0.117]		
Gr3 EFAL Word Recognition	23.121	25.928	22.729	-2.807*	0.392
	[0.900]	[1.190]	[1.429]		
Gr3 EFAL ORF at 60 seconds	27.255	30.634	27.647	-3.379*	-0.392
	[1.262]	[1.468]	[2.024]		
Gr3 EFAL ORF Comprehension	0.956	1.267	1.073	-0.311***	-0.117
	[0.059]	[0.087]	[0.100]		
Gr 3 EFAL Productive Vocabulary	3.120	3.509	3.345	-0.389***	-0.225
	[0.084]	[0.088]	[0.129]		
Gr 3 English Comprehension	0.863	1.186	1.012	-0.324***	-0.149*
	[0.041]	[0.073]	[0.078]		
Gr 3 HL Written Comprehension	2.441	2.286	2.072	0.155	0.370**
	[0.087]	[0.109]	[0.115]		
Gr 3 EFAL Written Comprehension	1.422	1.502	1.379	-0.080	0.043

	[0.049]	[0.076]	[0.080]		
Gr 3 Mathematics	1.422	1.502	1.379	-0.080	0.043
	[0.049]	[0.076]	[0.080]		
PCA: EFAL Language Proficiency	-0.000	0.387	0.197	-0.387***	-0.197*
	[0.055]	[0.079]	[0.100]		
PCA: EFAL Reading Proficiency	0.000	0.269	-0.021	-0.269**	0.021
	[0.078]	[0.106]	[0.131]		
PCA: HL Reading Proficiency	0.000	-0.133	-0.303	0.133	0.303**
	[0.076]	[0.082]	[0.107]		

The value displayed for t-tests are the differences in the means across the groups. Standard errors are clustered at the school level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 25: Comparing EFAL word recognition across the waves of data collection – Full sample

	End of Grade 1		End of Grade 2		End of Grade 3*	
	Decodable words	Sight words	Decodable words	Sight words	Decodable + Sight Combined	% Zero scores
Control	5	5.3	17	16.5	23.1	21%
On-site coaching	5.3	5.5	17.8	17.6	25.9	21%
Virtual coaching	4.6	4.7	16.5	16.3	22.7	24%

Figure 54: Improvements in oral reading fluency between Grade 2 and Grade 3 – Full sample

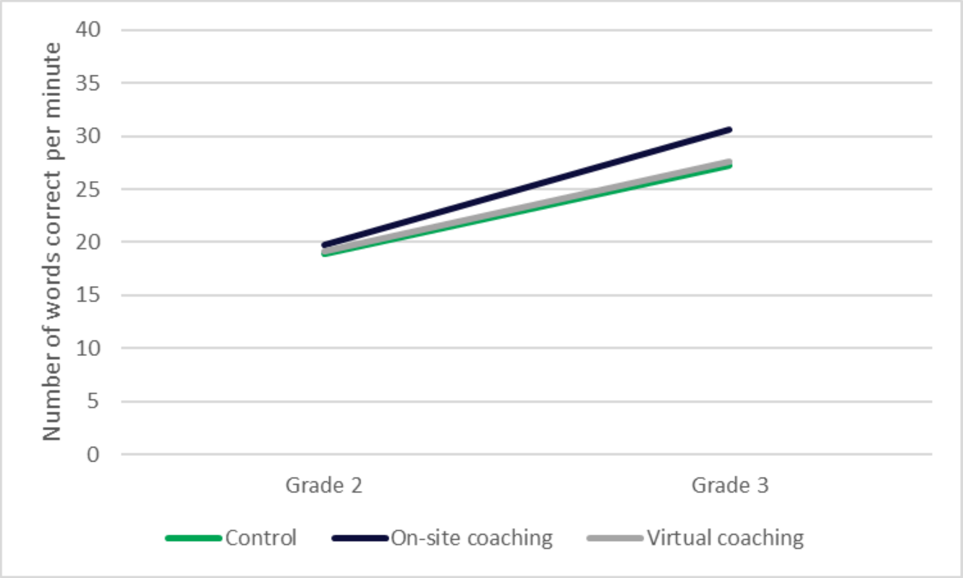


Figure 55: Performance distribution for EFAL word recognition and ORF - Full sample

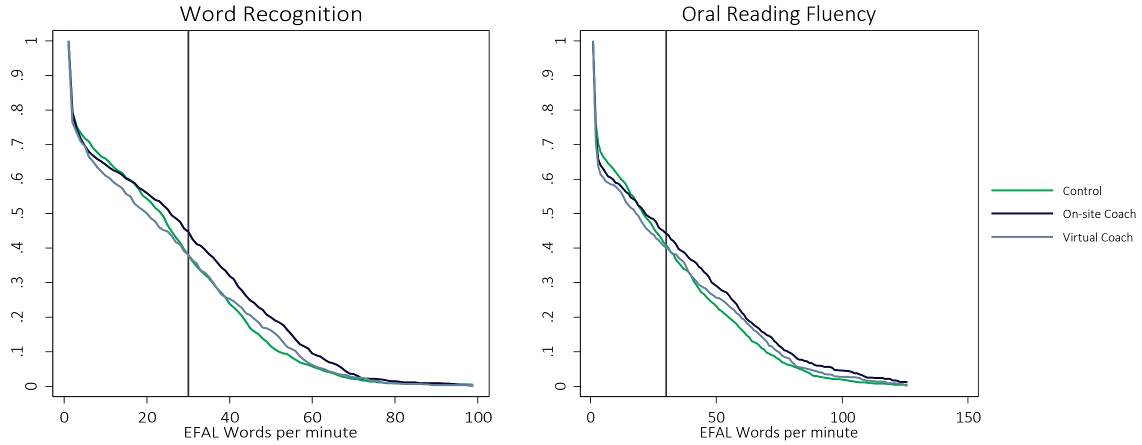


Figure 56: English reading and writing comprehension - Full sample

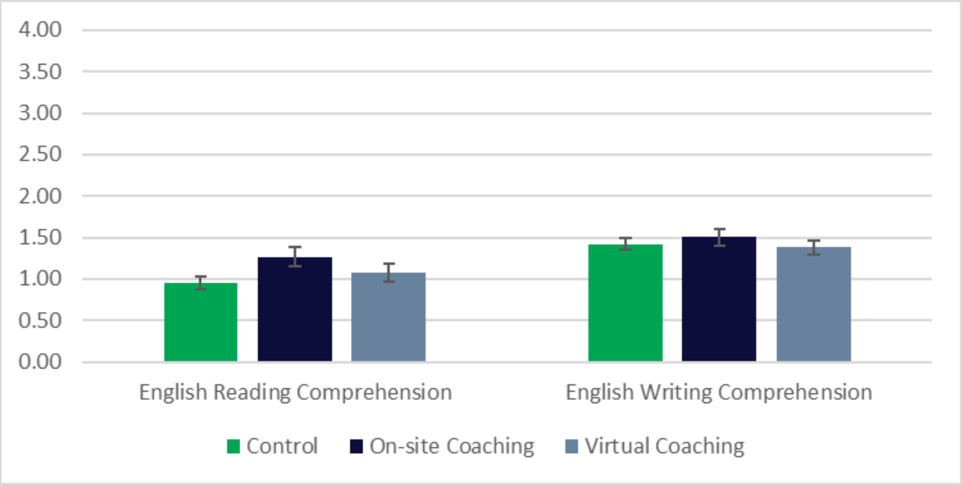


Figure 57: English vocabulary - Full sample

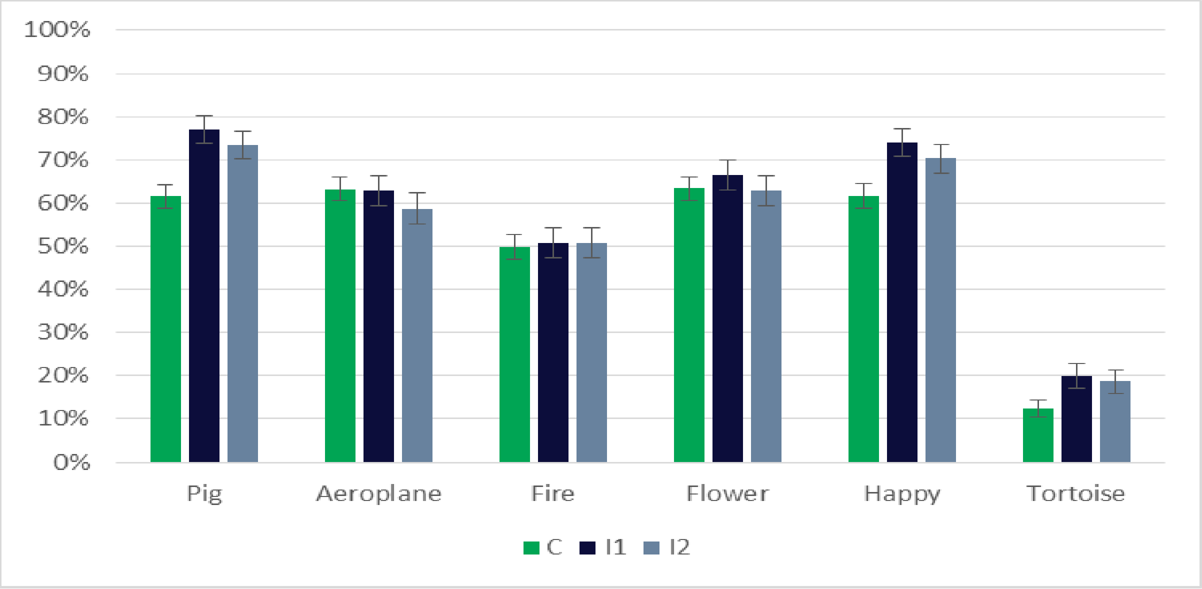


Figure 58: English listening comprehension score

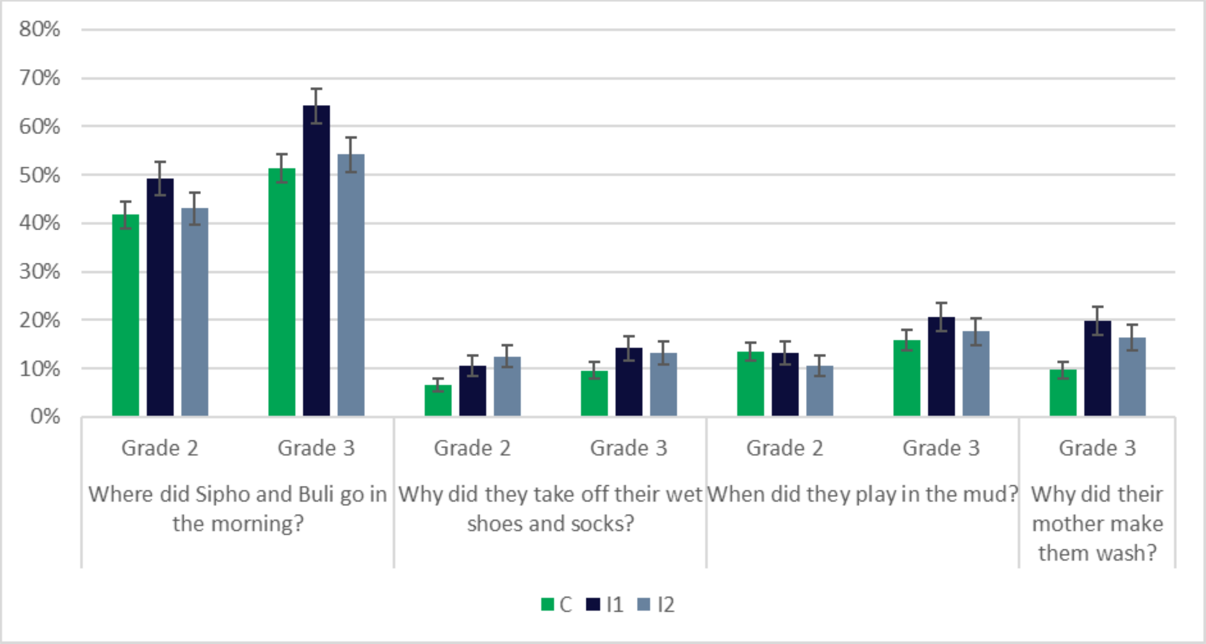


Table 26: Comparing HL letter-sound recognition - Full sample

	Start of Grade 1	End of Grade 1	End of Grade 2	End of Grade 3
Control	7	17.7	37.1	42.9
On-site coaching	6.8	16.7	37.8	48.1
Virtual coaching	7	15.1	34.8	41.7

Table 27: Comparing HL word reading – Full sample

	Word Reading End of Grade 1		ORF End of Grade 2		ORF End of Grade 3		
	Average words	% of Zero Scores	Average words	% of Zero Scores	Average words	% of Zero Scores	Mean Comp.
Control	5.5	45.7%	15.7	34.1%	23.1	21.5%	2.4
On-site coaching	4.7	49.1%	13.8	41.3%	21.5	26.5%	2.3
Virtual coaching	4.7	51.8%	13.5	43.6%	20.7	27.7%	2.1

Figure 59: Distribution of HL oral reading fluency – Full sample

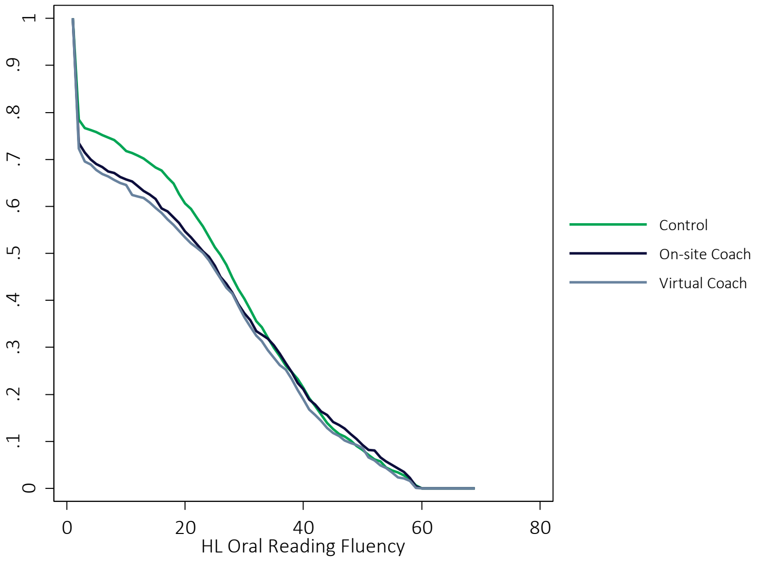


Table 28: Regression on the English learning outcomes - Grade 3 sample

VARIABLES	(1) Oral Prof.	(2) Decoding	(3) Vocab.	(4) L. Compr.	(5) Word Recog.	(6) Read. Fluency	(7) Read. Compr.	(8) W. Compr.
On-site coach	0.356***	0.179**	0.258***	0.351***	0.176***	0.122*	0.271***	0.104
	(0.078)	(0.072)	(0.064)	(0.084)	(0.067)	(0.071)	(0.085)	(0.075)
Virtual coach	0.149*	-0.015	0.086	0.170**	-0.031	0.016	0.106	-0.033
	(0.078)	(0.076)	(0.068)	(0.082)	(0.072)	(0.075)	(0.078)	(0.069)
Observations	2,148	2,109	2,148	2,148	2,148	2,148	2,148	2,109
R-squared	0.270	0.273	0.248	0.213	0.240	0.241	0.248	0.183
P-value	0.0223	0.0208	0.0185	0.0614	0.00786	0.187	0.0835	0.100
Control mean	0.139	0.170	0.130	0.108	0.159	0.152	0.126	0.161

* Controlling for learner gender, learner age, baseline scores, district, school quintile strata and fieldworkers. Only includes Grade 3 learners.

Figure 60: Effect of on-site coaching on English learning outcomes – Full sample

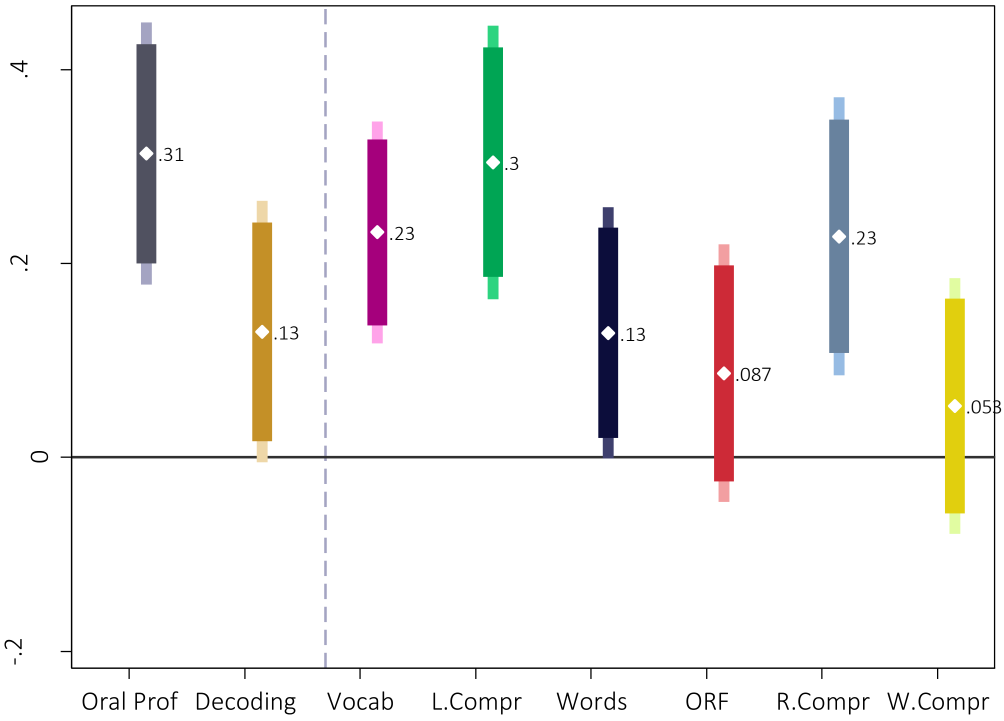


Figure 61: Effect of virtual coaching on English learning outcomes – Full sample

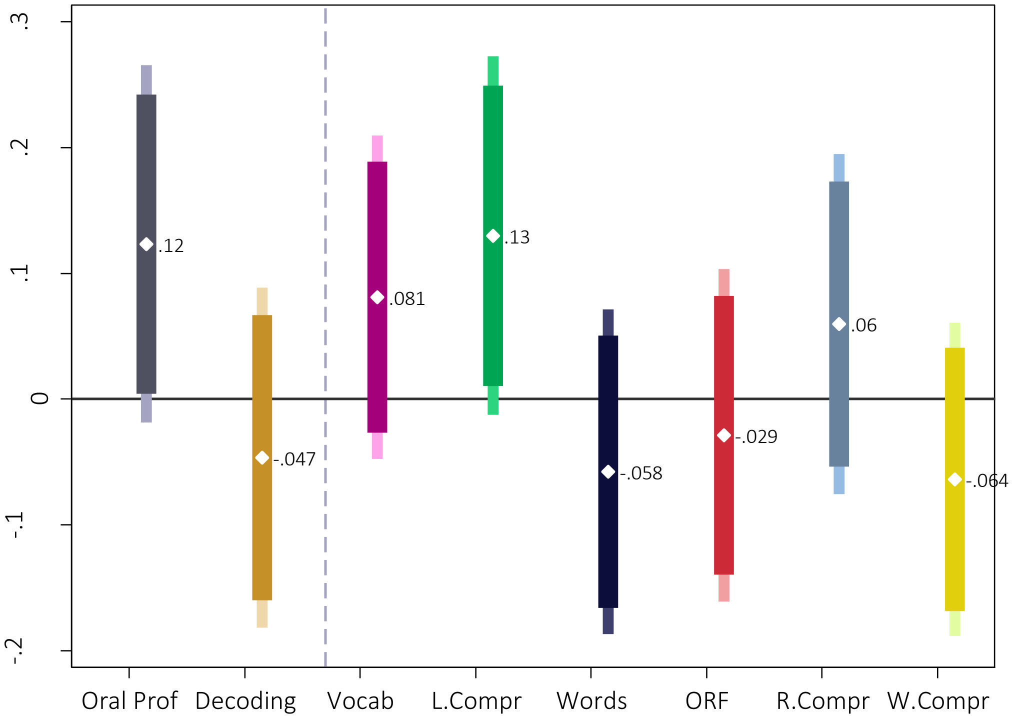


Figure 62: Learning gains in oral reading fluency between Gr 2 and Gr 3 – Full sample

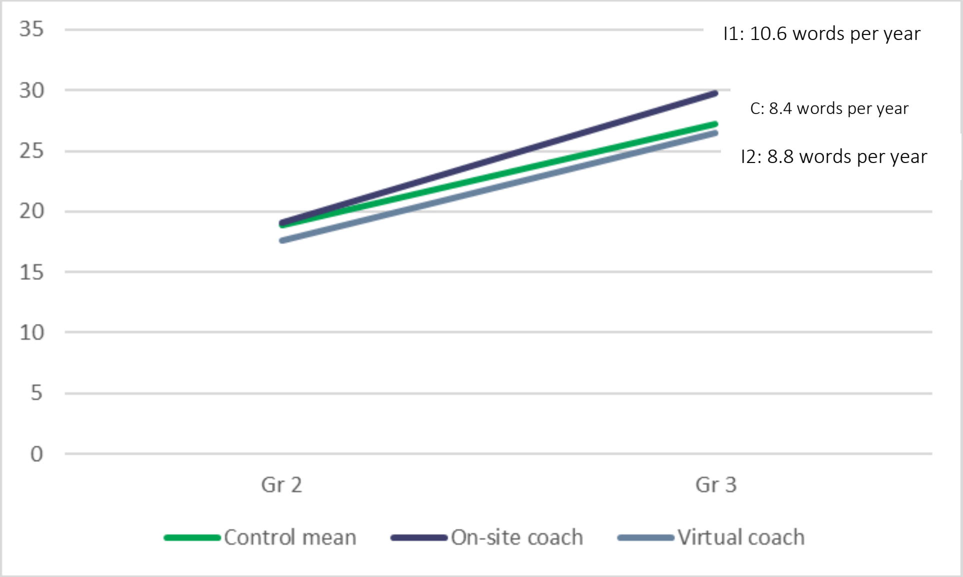


Figure 63: Learning gains in English listening comprehension between Gr 2 and Gr 3

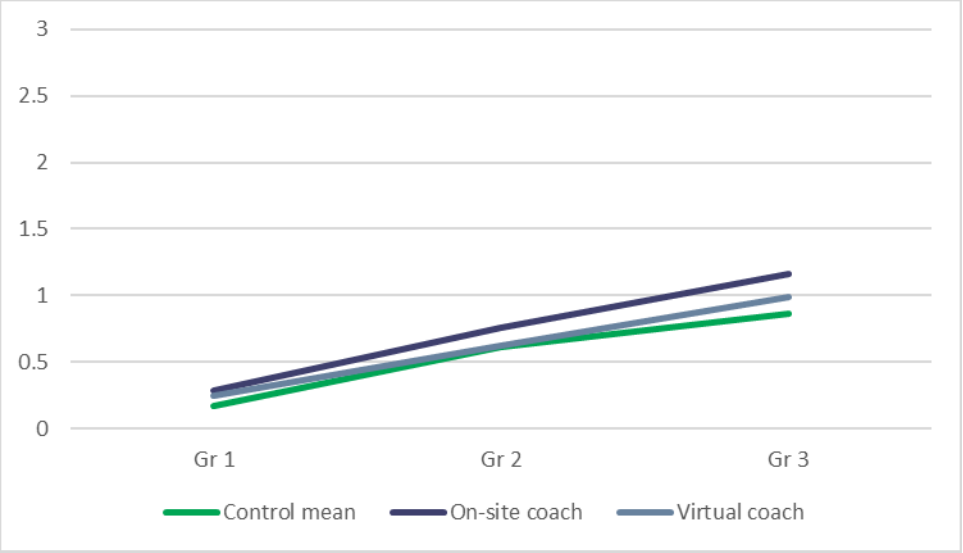


Figure 64: Girls performing better than boys in reading proficiency – full sample

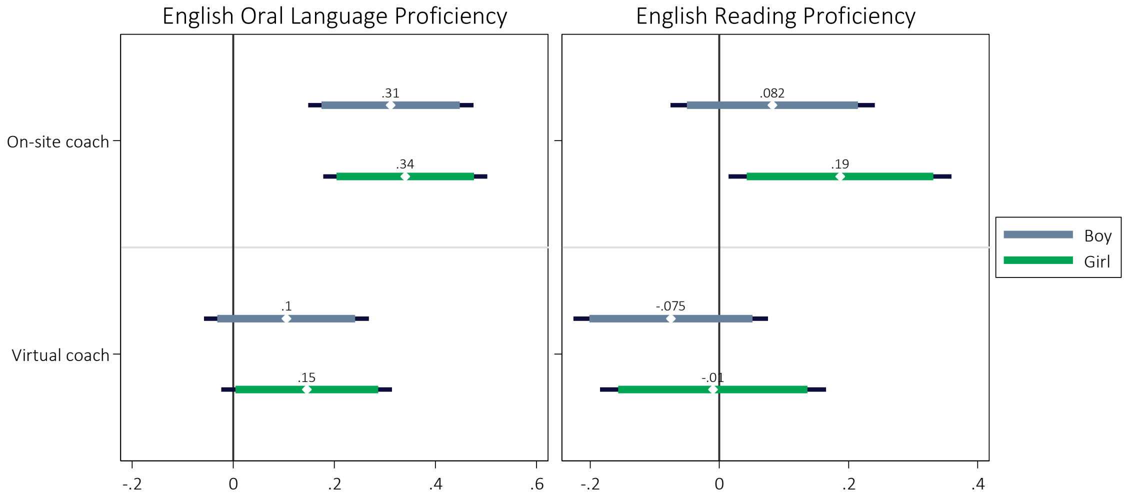


Figure 65: Girls benefitting more from the on-site coaching intervention in reading proficiency - Full sample

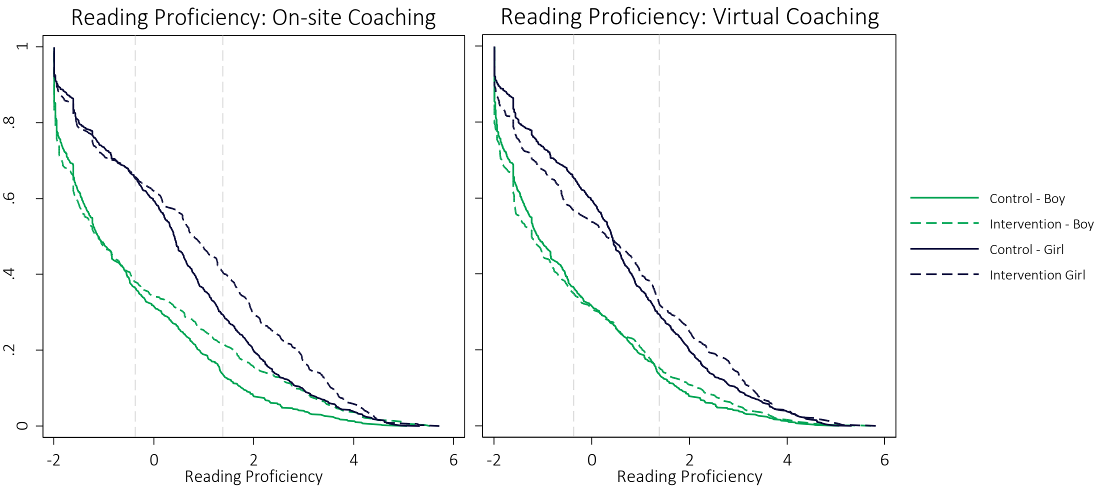


Figure 66: Learners in Siswati schools seems to be benefitting more from both interventions in language proficiency

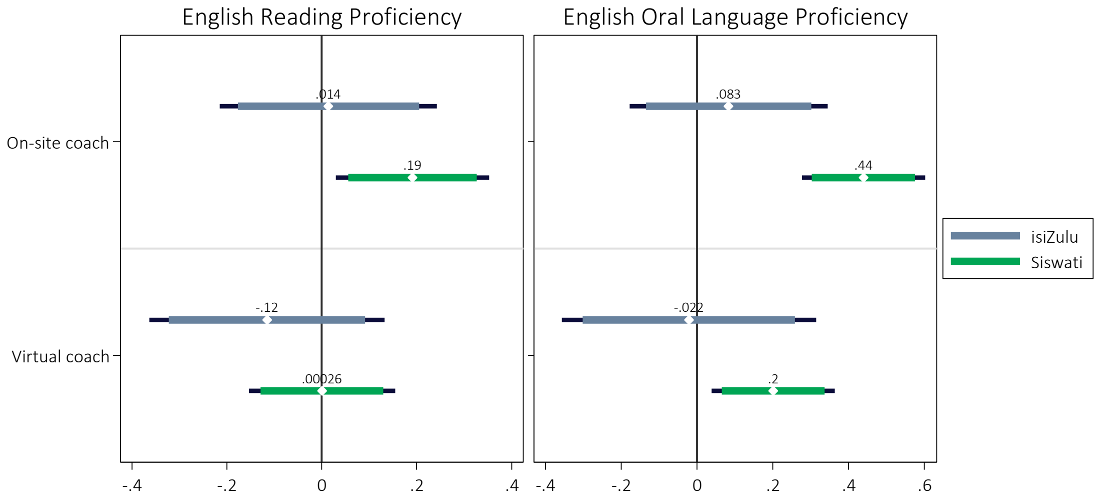


Figure 67: In reading proficiency, learners in Siswati schools benefitting more from the on-site coaching intervention

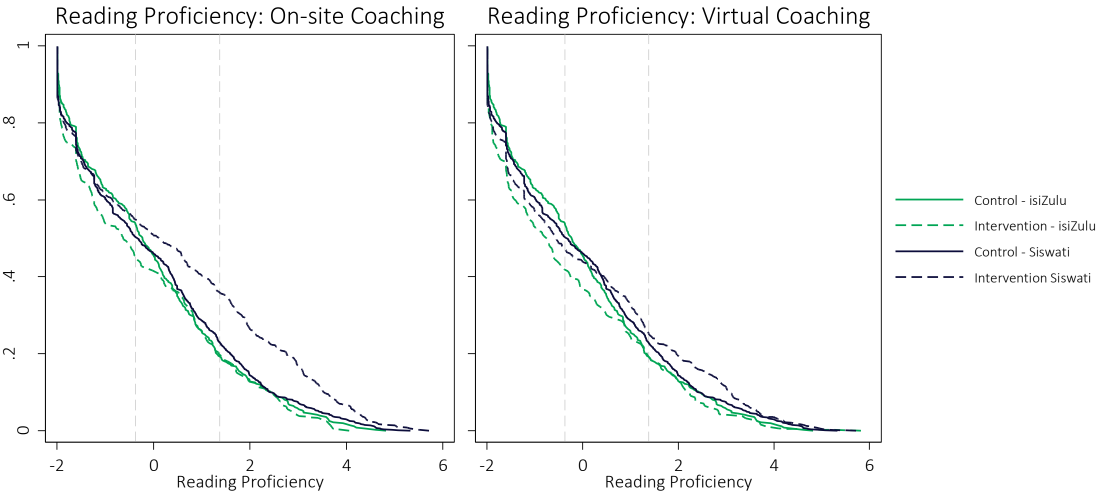


Figure 68: Difference between the intervention groups across the performance distribution – full sample

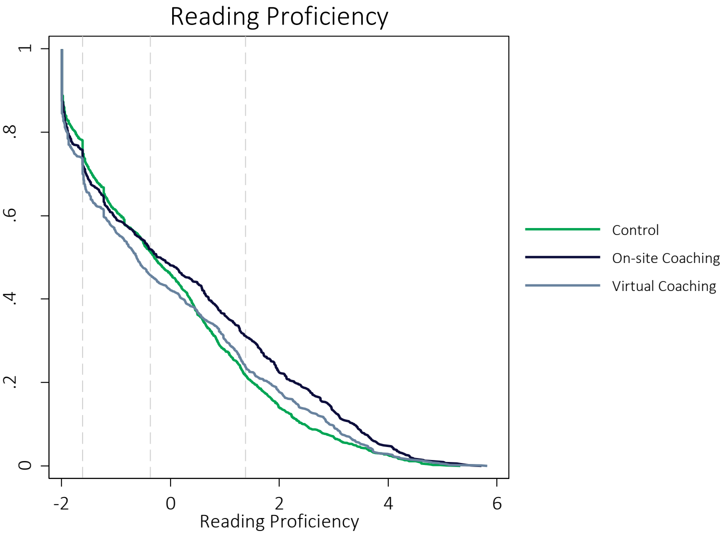


Figure 69: Effect of on-site coaching on HL and Maths – Full sample

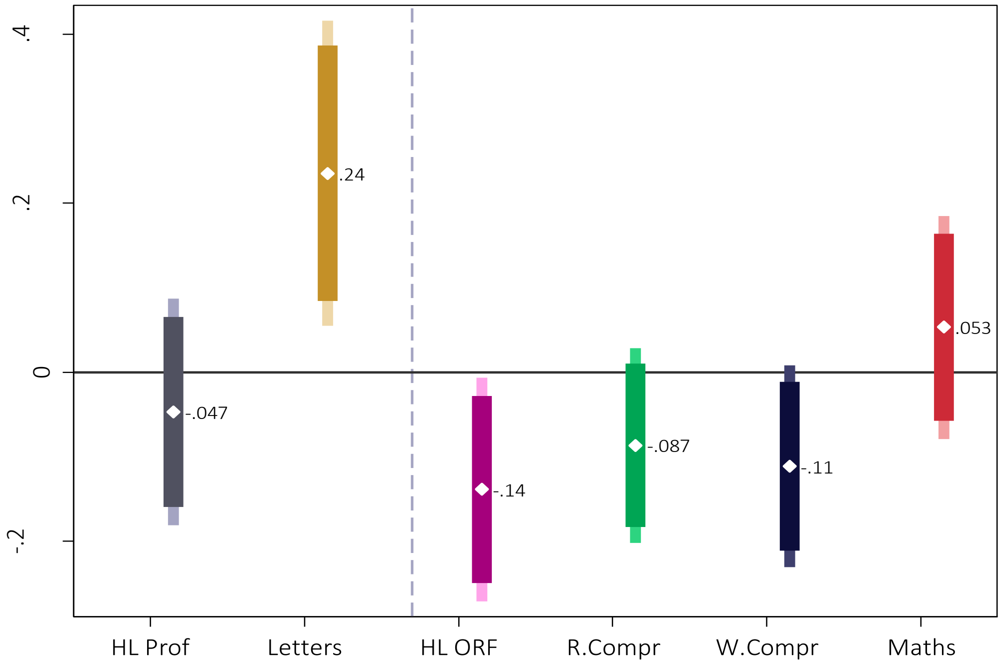


Figure 70: Effect of virtual coaching on HL and Maths – Full sample

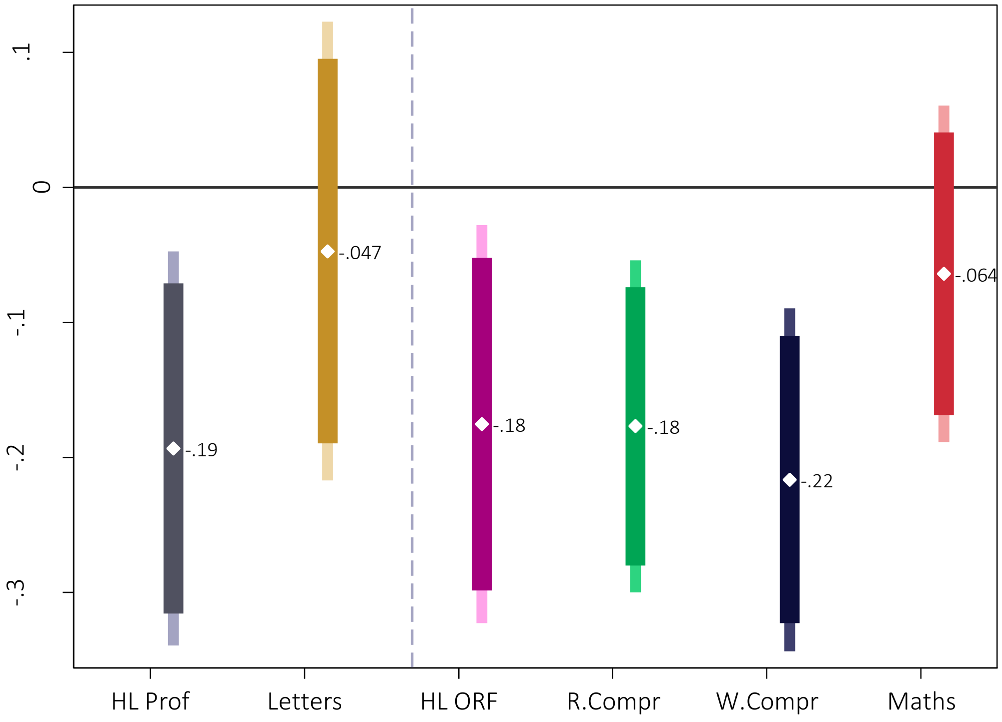


Table 29: Implementation Fidelity - index and components

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Implementation Fidelity Index	Did you receive training in EFAL in 2019?	Observed by a coach more often than once a year	Coach modelled a lesson more often than once a year	Have you ever received a compliment from a coach?	Have EFAL Readers	Number of EFAL Readers	EFAL Teaching Time
T1	1.367***	0.262***	0.562***	0.505***	0.528***	0.297***	1.293***	0.490**
	(0.122)	(0.058)	(0.062)	(0.068)	(0.062)	(0.053)	(0.223)	(0.241)
T2	0.947***	0.234***	0.160**	0.280***	0.161**	0.314***	1.802***	0.380
	(0.141)	(0.064)	(0.077)	(0.077)	(0.075)	(0.054)	(0.214)	(0.233)
Observations	296	292	296	296	296	279	296	280
R-squared	0.331	0.106	0.232	0.188	0.237	0.207	0.256	0.056
P-value	0.00309	0.586	1.06e-06	0.00708	3.51e-06	0.636	0.0181	0.663
Control mean	-0.708	0.639	0.252	0.244	0.319	0.683	1.733	6.787

Standard errors are clustered at school level; * for p<.1; ** for p<.05; *** for p<.01; Only controlling for stratification dummies

Table 30: Teacher practice - index and components

	(1)	(2)	(3)	(10)	(11)	(12)	(13)	(14)
VARIABLES	Teacher behaviour Index	Use EFAL lesplan provided by NGO	Do you use EFAL readers everyday?	Correct: Phonics sound	Correct: Group guided reading	Correct: Phonics lesson	Correct: Shared reading	Correct: Creative writing
T1	1.191***	0.338***	0.389***	0.351***	0.469***	0.209***	0.110	0.288***
	(0.143)	(0.067)	(0.069)	(0.073)	(0.067)	(0.072)	(0.067)	(0.071)
T2	1.043***	0.418***	0.329***	0.255***	0.240***	0.194**	0.156**	0.280***
	(0.135)	(0.065)	(0.073)	(0.074)	(0.079)	(0.077)	(0.067)	(0.076)
Observations	296	296	296	296	296	296	296	296
R-squared	0.296	0.184	0.161	0.141	0.172	0.084	0.066	0.092
P-value	0.326	0.194	0.434	0.239	0.00748	0.861	0.516	0.913
Control mean	-0.632	0.519	0.407	0.400	0.222	0.444	0.637	0.407

Table 31: Teacher practice - components continued

	(1)	(2)	(3)	(4)	(5)	(3)
VARIABLES	Quality of EFAL wall posters	Quality of EFAL Flashcards	Story books in the class?	EFAL: Number of written activities	EFAL: Number of pages with full sentence	EFAL: Number of pages with full paragraph
T1	0.428***	0.823***	0.775***	-1.643	-2.518*	-0.438
	(0.156)	(0.147)	(0.166)	(2.246)	(1.289)	(0.878)
T2	0.406***	0.525***	0.912***	-4.248**	-3.124***	-0.599
	(0.141)	(0.157)	(0.164)	(1.832)	(1.119)	(0.805)
Observations	292	292	292	282	277	274
R-squared	0.078	0.148	0.173	0.054	0.089	0.055
P-value	0.894	0.0582	0.398	0.247	0.629	0.881
Control mean	2.707	2.609	2.474	37.46	16.54	5.715

Standard errors are clustered at school level; * for p<.1; ** for p<.05; *** for p<.01; Only controlling for stratification dummies

Figure 71: Teaching skills acquired

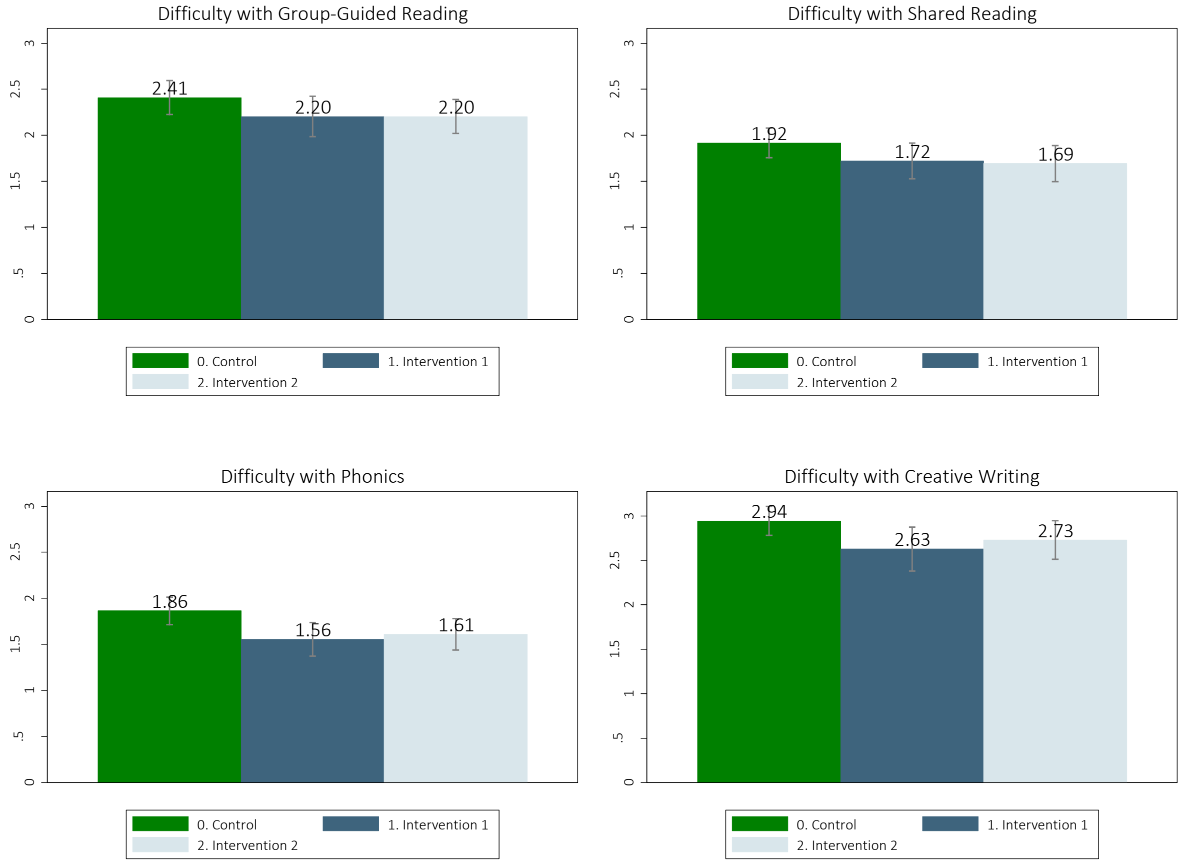


Table 32: Teaching skills acquired: Index and components

	(1)	(2)	(3)	(4)	(5)
	Skill			Difficult:	Difficult:
	acquisition	Difficult:	Difficult:	Shared	Creative
VARIABLES	Index	GGR	Phonics	Reading	Writing
T1	-0.370**	-0.204	-0.308**	-0.181	-0.300**
	(0.170)	(0.140)	(0.119)	(0.128)	(0.145)
T2	-0.367**	-0.212	-0.258**	-0.229*	-0.226
	(0.147)	(0.135)	(0.118)	(0.132)	(0.139)
Observations	294	290	293	292	290
R-squared	0.079	0.066	0.047	0.044	0.063
P-value	0.986	0.955	0.694	0.734	0.646
Control mean	0.170	2.376	1.828	1.872	2.933

Standard errors are clustered at school level; * for p<.1; ** for p<.05; *** for p<.01; Only controlling for stratification dummies

Figure 72: Support received from the school management team

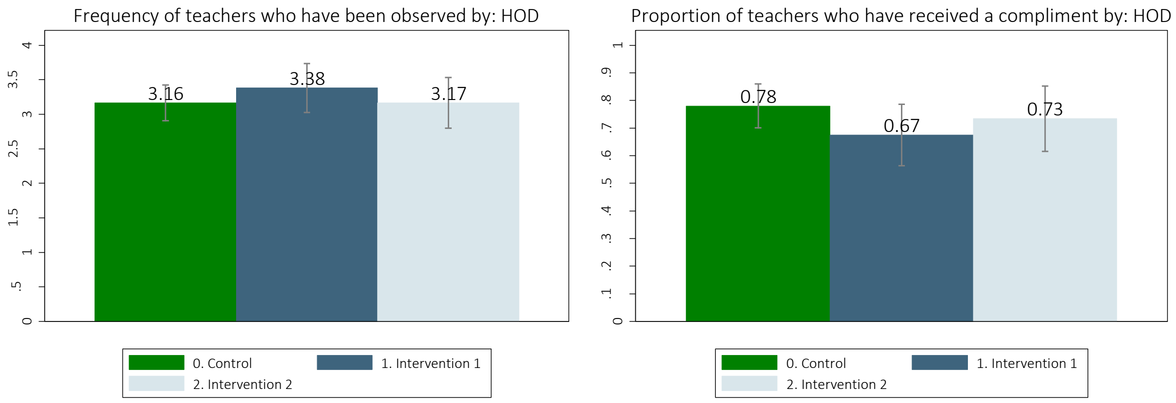


Table 33: SMT support received: Index and components

VARIABLES	(1) SMT Support Index	(2) Meeting as CoP	(3) Principal Observed	(4) HOD Observed	(5) Principal Modelled	(6) HOD Modelled	(7) Principal Compl.	(8) HOD Compl.
T1	0.109	0.106	0.188	0.347*	0.001	-0.187	0.145	-0.006
	(0.152)	(0.076)	(0.193)	(0.209)	(0.173)	(0.218)	(0.221)	(0.272)
T2	0.009	0.012	-0.106	0.077	-0.144	0.080	0.098	0.073
	(0.161)	(0.081)	(0.189)	(0.217)	(0.192)	(0.238)	(0.247)	(0.245)
Observations	292	292	282	267	283	267	282	271
R-squared	0.071	0.022	0.100	0.078	0.114	0.066	0.052	0.013
P-value	0.575	0.286	0.155	0.263	0.465	0.302	0.865	0.801
Control mean	-0.0405	0.617	2.654	3.143	1.891	2.686	2.535	3.098

Standard errors are clustered at school level; * for p<.1; ** for p<.05; *** for p<.01; Only controlling for stratification dummies

Table 34: Correlation between retest and initial scores

	Correlation
Letter Naming	0.8086
HL ORF	0.9252
HL ORF Comp	0.8498
EFAL ORF	0.9213
EFAL ORF Comp	0.7959

Table 35: Potential spillover effects - variety of activities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Singing songs/ rhymes	Vocab development occurred	Shared Reading occurred	Shared Reading of short passages occurred	Group-guided reading occurred	Individual/ Paired reading occurred	Writing activity occurred
T1	-0.018	-0.295***	0.091	0.070	0.236	0.173	-0.024
	(0.116)	(0.103)	(0.150)	(0.142)	(0.150)	(0.137)	(0.139)
T2	0.060	-0.102	0.013	0.219	0.058	0.122	0.228*
	(0.128)	(0.087)	(0.139)	(0.160)	(0.146)	(0.139)	(0.130)
Observations	957	957	937	937	938	957	957
R-squared	0.070	0.436	0.214	0.150	0.187	0.140	0.254
Control mean	0.145	0.937	0.548	0.208	0.211	0.178	0.620

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Only controlling for stratification dummies