RISE Working Paper 20/050
January 2021

Can Virtual Replace In-person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning

Jacobus Cilliers, Brahm Fleisch, Janeli Kotzé, Nompumelelo Mohohlwane, Stephen Taylor, and Tshegofatso Thulare

Abstract

Virtual communication holds the promise of enabling low-cost professional development at scale, but the benefits of in-person interaction might be difficult to replicate. We report on an experiment in South Africa comparing on-site with virtual coaching of public primary school teachers. After three years, on-site coaching improved students' English oral language and reading proficiency (0.31 and 0.13 SD, respectively). Virtual coaching had a smaller impact on English oral language proficiency (0.12 SD), no impact on English reading proficiency, and an unintended negative effect on home language literacy. Classroom observations show that on-site coaching improved teaching practices, and virtual coaching led to larger crowding-out of home language teaching time. Implementation and survey data suggest technology itself was not a barrier to implementation, but rather that in-person contact enabled more accountability and support.

The original version of this paper was first published in September 2020.

JEL Classification: C93, I21, I25, I28, O15

Can Virtual Replace In-person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning

Jacobus Cilliers
McCourt School of Public Policy,
Georgetown University
ejc93@georgetown.edu

Nompumelelo Mohohlwane
Department of Basic Education, Government of South Africa

Brahm Fleisch
University of Witwatersrand's School of Education, South Africa

Stephen Taylor
Department of Basic Education, Government of South Africa

Janeli Kotzé
Department of Basic Education, Government of South Africa

Tshegofatso Thulare
Department of Basic Education, Government of South Africa

Acknowledgements:

We are grateful for USAID's financial support, and useful feedback from Jishnu Das, David Evans, and Abhijeet Singh. The study was registered with the AEA Trial Registry: https://doi.org/10.1257/rct.5148-1.0.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Foreign, Commonwealth and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Cilliers et al. 2021. Can Virtual Replace In-person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning. RISE Working Paper Series. 20/050. https://doi.org/10.35489/BSG-RISE-WP_2020/050.

Use and dissemination of this working paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The findings, interpretations, and conclusions expressed in RISE Working Papers are entirely those of the author(s) and do not necessarily represent those of the RISE Programme, our funders, or the authors' respective organisations. Copyright for RISE Working Papers remains with the author(s).

1 Introduction

Virtual communication holds the promise of enabling low-cost professional development at scale, but the benefits of in-person interaction might be difficult to replicate. On the one hand, an expert instructor or coach could reach a far wider audience virtually, which reduces transport and salary costs, and could overcome binding human resource constraints. On the other hand, employees 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 employee and the instructor or coach, which allows her to be vulnerable and discuss ways to improve her knowledge or performance. A lack of face-to-face engagement could also reduce accountability.

These trade-offs are particularly evident in the context of professional development support for teachers in developing countries. There is an urgent need to improve teaching capacity, given low levels of learning, highly challenging teaching environments, and weak teacher mastery of content and pedagogical skills (Bold et al., 2017). Teacher professional development programs are ubiquitous —most of the approximately 90 million teachers in the world receive some kind of in-service teacher training on an annual basis— but governments typically implement low-cost teacher training models that are not grounded in evidence (Popova et al., 2016). A proven, effective way for teacher professional development is the use of pedagogical coaches, but such programs can be expensive. Given this challenge, virtual coaching provides on opportunity to deliver high-quality coaching at scale, but it might also be less effective in environments with limited technological penetration and weak accountability systems.

Can virtual replace on-site coaching? We address this question in the context of teacher professional development for teaching English as a Second Language (ESL) in the early grades in South Africa. Working with South Africa's Department of Basic Education (DBE), we randomly assigned 100 schools to receive either virtual or on-site coaching support, and another 80 schools to the control, where teachers could still receive business-as-usual professional development support provided by government.[1] In both programs teachers received the same learning materials and training

[1]The DBE requested that no other interventions targeting the teaching of ESL are implemented in the control schools, but allowed teachers from both the intervention and control schools to attend any other teacher training that is organised and presented by the district or the province.

at the start of the program, and the curriculum and content of lesson plans were the same and fully aligned with government curriculum. However, the on-site coaching intervention differed from the virtual coaching in two important dimensions. First, teachers in the on-site program would receive in-classroom visits by a coach, whereas teachers in the other program would interact virtually with a coach through phone calls, regular text messages, WhatsApp groups, and participation in competitions. Second, the format of the daily lesson plans was paper-based in the on-site coaching intervention but was on an electronic tablet in the virtual coaching intervention.

These programs were implemented over a period of three years, targeting the teachers assigned to a different grade each year (grade one teachers in the first year, grade two teachers in the second year, etc). We randomly sampled and assessed 20 grade one students per school before the start of the program in February 2017. We then tracked the same cohort of students over a period of three years, starting in February 2017 when they entered grade one, and ending in November 2019. At the end of every school year these students were assessed and their teachers surveyed. We also performed classroom observations in a sub-set of 53 schools at the end of the third year.

We highlight five main findings. First, the on-site coaching intervention was more effective at improving English reading proficiency, relative to virtual coaching. After three years, on-site coaching improved both English oral language proficiency (0.31 standard deviations) and English reading proficiency (0.13 standard deviations). In contrast, virtual coaching was far less effective at improving English oral language proficiency (0.12 standard deviations), and had no statistically detectable impact on reading proficiency skills. Moreover, quantile regressions reveal that for both programs the best-performing students experienced the largest gains in reading proficiency. The on-site coaching program is about 23% more expensive than the virtual coaching program, but the cost-effectiveness analysis shows that it is still more cost-effective.

Second, the pattern of effectiveness changed over time. By the end of the first year the on-site and virtual programs were equally effective at improving English oral language proficiency —by 0.55 and 0.52 standard deviations, respectively— and neither had an impact on reading proficiency.[2] These dynamic effects reflect the fact that the ESL curriculum focuses exclusively on developing oral language skills in the

[2]Some of the first-year results were reported by Kotze et al. (2019).

first year, with a shift towards teaching reading skills in the second and third years. The pedagogical techniques required for teaching reading skills are more technically challenging, relative to teaching oral language.

Third, classroom observations reveal that the on-site coaching induced larger gains in teacher productivity, relative to virtual coaching, especially for teaching practices that develop reading proficiency. 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 students, such as group-guided reading, were better and more frequently implemented by teachers who had received on-site coaching. Moreover, mediation analysis provides suggestive evidence that the improvements in group-guided reading in the on-site arm was a key ingredient for improving reading proficiency skills (Acharya et al., 2016).

Fourth, virtual coaching *reduced* home language reading proficiency by 0.19 standard deviations and caused a reallocation of time inputs away from home language (HL) to ESL instruction. Time usage data reveal that teachers in both programs dedicate less time to HL instruction, but this reduction is more pronounced in the virtual arm.

Fifth, we are able to rule out differences in fidelity of program implementation or technology failure as the reason for the virtual coaching program being less effective than on-site coaching program, and therefore conclude that the critical difference was the nature of the coaching interaction. The same service provider implemented both programs, and the quality of implementation was equally high for all three years of the study. Moreover, tablet usage data show that technology itself was not a barrier to program implementation, since almost all the teachers in the virtual arm used tablets and accessed the lesson plans. Rather, the pattern of tablet usage —which was better earlier in the term, and highest in the week in which teachers were expected to submit assessment results— suggest that the binding constraint was teacher motivation or their ability to keep pace with the curriculum, rather than the technology itself. Consistent with this interpretation, teachers in the on-site coaching intervention were far more likely than teachers in the virtual coaching intervention to mention the coach as someone who holds them accountable and provides pedagogical support.

This study contributes to two strands of literature. First, in terms of teacher professional development, a growing body of research from developing countries has

demonstrated the important role that pedagogical coaches can play in improving student learning (Kraft et al., 2018) especially when combined with a carefully planned curriculum (often with daily lesson plans), and additional learning aids (Cilliers et al., 2019; Eble et al., 2020; Evans and Popova, 2016; Piper et al., 2014; Snilstveit et al., 2014). This study contributes to this literature by testing for a more cost-effective modality of delivery. This is important, since there are concerns about the scalability of coaching programs, as well as the effectiveness of less expensive variants (Kerwin and Thornton, 2020).

Second, it contributes to the literature on the use of information technology in improving education outcomes. Previous studies have found that computer-assisted instruction can be highly effective at improving learning, particularly if it complements rather than substitutes teaching time, and is aligned with student ability (Banerjee et al., 2007; Beg et al., 2019; Muralidharan et al., 2019). But few studies have used experimental or quasi-experimental designs to examine the less expensive role that technology can play through improving teacher capacity in developing countries (examples include Piper et al. (2016) and Bruns et al. (2017)), and none experimentally compare virtual with on-site pedagogical support.[3]

Our findings are in contrast to evidence from the United States on the relative effectiveness of on-site versus virtual coaching. In a meta-analysis of evaluations of coaching interventions, Kraft et al. (2018) found no statistically discernible difference in effect size between in-person and virtual coaching, although they note limited statistical power to rule out "even moderately sized differences". Powell et al. (2010) experimentally compared virtual with on-site coaching of pre-K teachers, and found that after one semester the programs were equally effective at improving oral language proficiency. This is consistent with the first-year results of this evaluation. The fact that results from this evaluation changed when assessing reading skills after three years of exposure to the program, highlights the importance of longer-term studies that assess different domains of student learning over time. As McEwan (2015) notes, most studies on education interventions show impacts after just one year (or less).

[3]Piper et al. (2016) found that giving tablets to teachers did not increase the effectiveness of an existing teacher professional development program in Kenya. Bruns et al. (2017) found that online coaching in Brazil had a modest improvement (0.04 to 0.08 SD) in student learning.

2 Sample, program description, and experimental design

2.1 Background and sample

The study is set in two districts in the Mpumalanga province in South Africa. Mpumalanga is a mid- to low- performing province in terms of education performance, and is one of the poorest provinces in the country. In the 2019 national end-of-high-school examinations, Mpumalanga ranked sixth out of the nine provinces. According to the 2016 General Household Survey, 28.4 percent of students attending schools in Mpumalanga fell below the food poverty line (monthly per capita income is below R442.00 ($24)). The two districts were chosen because they are relatively linguistically homogeneous: the majority of schools either have isiZulu or Siswati as the language of instruction.

As in many developing countries today, there is a growing awareness that the South African education system is producing alarmingly low levels of learning, especially in early grades (World Bank, 2018). Despite improvements in South Africa's performance in international assessments of literacy and numeracy over the past two decades, the average level of performance is still extremely low and is also highly unequal. A nationally representative assessment in 2016 found that 78 percent of South African grade four students did not reach the minimum literacy benchmark (Howie et al., 2017). This number was 83 percent in Mpumalanga. Moreover, studies have found that primary school classrooms are mainly characterised by a lack of print material, a lack of opportunities for reading and writing and weak instructional practices (Taylor, 2007). The EGRS interventions were designed to address these challenges.

South Africa is also similar to many developing countries in its linguistic diversity, with eleven official languages, but with English as the dominant language used in post-school education and spoken in commerce. As a result, the language policy balances the need for children to learn to read and write in a language they understand, with the need to develop proficiency in English. In practice, most children in South Africa learn in their home language as the main language of instruction during grades 1 to 3 and then experience a transition to English as the language of instruction from grade 4 onward.[4]

[4]Schools can either transition to Afrikaans or English, but the majority of schools transition to

To ameliorate the language transition learners face in grade 4, English is introduced as an additional language from grade one. According to the national curriculum, students should be taught oral language proficiency skills during the English lessons in the first grade, and decoding (i.e. reading skills) is only introduced in the English lesson from the second half of the second grade.[5] In the third grade, both oral language proficiency and reading proficiency skills are consolidated and students should be able to read for meaning by the end of the year.

2.2 Program description and experimental design

We evaluate the impact of two interventions aimed at improving teachers' enactment of the official English as a Second Language (ESL) curriculum in grades one to three.[6] Both interventions consist of three inter-related components: (1) detailed lesson plans, (2) integrated learning and teaching support materials, such as graded reading booklets, and (3) instructional coaching and training by a specialist reading coach. But they differ in the mode of coaching —virtual versus in-person— and the medium of the lesson plans— tablet or paper-based.

The content and support materials provided were the same in both interventions and were fully aligned to the official ESL curriculum. The lesson plans, following the curriculum guidelines, are explicit about the required weekly frequency of implementing different teaching activities (see Table A.1). In the first grade, teachers did phonics and phonemic awareness as well as shared reading activities with the class more frequently, as these activities focus on familiarising students with the new language. Group-guided reading (GGR) – an activity which requires a teacher to listen to a different group of five to eight students reading individually – was introduced in the second grade and should be implemented every day. The lesson plans also require that the teachers dedicate fours hours to teaching ESL and seven hours to teaching HL. As per the official curriculum guidelines, schools can choose between a 4:7 or a

English

[5] In contrast, the Home Language (HL) subject introduces reading skills in grade one.

[6] The study builds on and complements a previous early grade reading study (EGRS I) that targeted Home Language literacy in South Africa, which found that on-site coaching was more cost-effective at improving reading, compared to a traditional teacher training program in which teachers meet at a central location to receive training, but there were concerns about the scalablity of the program. In collaboration with the Department of Basic Education, this study (EGRS II) was developed with the question of cost and scalability in mind.

3:8 breakdown of hours dedicated to teaching ESL vis-a-vis HL.

The main difference between the two treatments was in the delivery model of the lesson plans and the coaching support (table A.2 provides a summary of the differences between the two interventions). In the first intervention, which we refer to as the *on-site* treatment arm, the teachers received a paper-based version of the lesson plans and benefited from regular on-site coaching with a specialised reading coach that visited the teachers in their classrooms. Coaches were required to visit each teacher 12 times a year. Figure A.1, panel (a), shows that teachers in the on-site coaching arm received between 5 to 25 visits in 2019, with the average teacher having received about 14 visits in the year. During these visits, coaches modelled, supported and evaluated teachers' practices and monitored implementation fidelity.

In the second intervention, which we refer to as the *virtual* treatment arm, the teachers received a tablet with an electronic version of the lesson plans, and they were supported by a virtual coach who called the teachers on a regular basis and sent weekly reminders and teaching tips through WhatsApp. The coach called every teacher at the start of the term, and followed up every two weeks if she felt that the teacher required additional support. Figure A.1, panel (b), shows teachers in the virtual arm received between 7 and 18 calls in 2019, with a mean number of 10 calls. The coach also received calls from the teachers and answered questions over WhatsApp on an ongoing basis.[7] In addition to the lesson plan, the tablets include additional electronic resources such as short training videos, sound clips of the phonics sounds, songs and rhymes, and examples of students' work.[8] The content was updated quarterly and designed to work offline; connectivity was therefore not a barrier for daily usage. Figure A.2 shows the distribution of time spent engaging with the tablet in the third term of the final year of the program: the average teacher spent 12.7 hours a term accessing content on the tablet.

The virtual coach also introduced small competitions around specific themes. Teachers were required to submit either videos or photos of their teaching for the competitions. The coach would then choose the best teacher in each of the teacher groups who was awarded with a small amount of airtime. The competitions were

[7] We unfortunately do not have data on the number of times that the teachers called or messaged the coach.

[8] A majority of the training videos were filmed in the classrooms in the evaluation sample. Therefore, teachers would see the methodologies enacted by teachers like themselves in classrooms that look similar to their own.

intended to give the virtual coach a way to observe actual teaching practice, thus enabling her to provide more targeted feedback. The competitions also helped teachers to see what other teachers in similar contexts were doing, thereby fulfilling the role of a virtual community of practice. Figure A.3 shows that participation in these competitions was variable: 78 percent of teachers participated in the competition at least once, but only 23 percent participated in every competition.

Teachers from both treatments received training at the start of each term. The first training session was residential training and entailed two days of training for the on-site treatment and three days of training for the virtual treatment, 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 teachers in the virtual treatment. The trainers rotated so that once during the year, all of the teachers in this intervention would be trained by the virtual coach once. School management team (SMT) members were also invited to attend the training, and a separate session was held to encourage and equip them to provide more regular support to the teachers in the intervention.[9] To reinforce this support, the virtual coach also communicated regularly with the SMTs over the duration of the year, and 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.

Figure A.4 shows that the attendance rates of teachers at the training sessions were very high (on average 98% attendance) with no difference in attendance between the treatment arms. In the case where teachers from either treatment arm did not manage to attend the training session, the on-site coaches organised a catch-up session to make sure that the teachers had the new materials and understood the instructional practices which were covered during the training. The attendance of SMTs at the training was not compulsory and was therefore much lower, and decreased over time. It is interesting to note that the attendance of SMTs from the virtual coaching schools was significantly lower than the attendance of SMTs from the on-site coaching schools.

The interventions were implemented with grade one teachers in 2017, grade two teachers in 2018 and grade three teachers in 2019, thereby following the same cohort

[9]The SMT in a school consists of the school principal, deputy principal and heads of departments (HODs), and are responsible for providing instructional leadership and support to teachers

of students. About 7,600 students benefited from the interventions for the three year period. Teachers typically teach the same grade every year, so a different group of teachers were exposed to the program each year.

For purpose of the evaluation, we randomly selected 180 public primary schools out of a population of schools that are non-fee paying public schools[10], whose primary language of instruction is Siswati or isiZulu, and whose grade one enrollment is between 30 and 160. [11] We then created 10 strata of similar schools, based on school size, socio-economic status and previous performance in a standardized national exam, and randomly assigned five schools to each intervention group and eight to the control group. Thus we randomly assigned 50 schools to each intervention and 80 to the control. Furthermore, within each school we randomly selected 20 grade one students, and tracked these students over a period of three years. One school in the See Figure A.5 for a summary of sample selection, take-up and attrition across all evaluation arms and waves of data collection.

3 Potential mechanisms

Broadly speaking, coaches can play three roles. The first is providing pedagogical support, where the coach gives targeted feedback to teachers on their instructional practices. The second role is one of accountability, where the coach monitors teachers' curriculum coverage to ensure that teaching is happening as required by the curriculum. The final is one of a confidante, where the coach builds a trust relationship with teachers that would emotionally prepare teachers for changing their instructional practices.[12]

The virtual coach faced three challenges that the on-site coach did not have in per-

[10] In South Africa public schools are classified into so-called "poverty quintiles", which are not exactly equally sized. The bottom three quintiles of schools do not charge fees and do receive a higher per-student government subsidy. These schools serve about 70 per cent of South African children.

[11] We excluded the smallest schools, because they were most likely to have multi-grade classes for which grade-specific lesson plans would not work; and we excluded the largest schools because of cost considerations

[12] Qualitative work conducted by the research team noted that past experience often conditions teachers to expect negative feedback from observers, but without any guidance to meet the expectations of the observer. In order for a teacher to move towards the openness and vulnerability needed for real behaviour change to take place, teachers need to be in an environment of trust and have clear and attainable expectations.

forming these functions, all linked to the lack of in-person classroom visits. Firstly, communication was limited to phone calls and text messages, making it harder to build a relationship of trust. 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 virtual coach could not observe classroom practice directly, and was therefore limited in the ability to provide targeted pedagogical support. Finally, accountability may have been weaker, since the monitoring of teaching activities was again dependent on information volunteered by teachers and could not be verified through direct observation. Efforts were made to mitigate these challenges such as the competitions where teachers sent videos of their teaching activities, creating the opportunity for each teacher to physically meet the virtual coach at least once at the centralized training sessions and engaging with the SMT to promote accountability.

4 Data and empirical strategy

4.1 Data collection

We conducted four rounds of data collection: once at the beginning of the first year of the program when the students started grade one (February 2017), and again at the end of each academic year (November, 2017, 2018, and 2019). During these rounds of data collection we conducted student assessments on the same panel of students, administered teacher surveys to those concurrently exposed to the intervention (grade one teachers the first year, grade two teachers the second year, etc.) and performed document and classroom inspections. We also surveyed the head teachers. (See Figure A.6 for a schematic summary of the timeline of the intervention and data collection.)

The components of the student assessments were adjusted each year to assess the oral language and decoding skills expected by the end of each year. At the end of the third grade we administered both an oral and a written assessment to the students in the sample. The student assessments were designed to evaluate students' language and literacy abilities at the end of each grade, but were not designed to necessarily benchmark student performance against curriculum requirements. Given this focus, the assessments included the EGRA-type tasks, and care was taken to minimize a floor effect. The oral assessments were administered by fieldworkers in an one-on-one

setting with the sampled students, whereas the written assessments were administered in a group setting. In the final wave of data collection, the oral assessment included eight tasks assessing oral and reading proficiency in HL and ESL. These tasks included HL letter recognition, HL oral reading fluency and comprehension, ESL expressive vocabulary, ESL listening comprehension, ESL word reading and ESL oral reading fluency and comprehension. A further written assessment was conducted with the students to assess their written comprehension abilities in both languages, as well as their basic mathematics skills. Table A.3 provides a summary of the different components of student assessment administered in the different years.

As specified in our pre-analysis plan, we evaluate the overall impact of the interventions using two indices that are based on the two language constructs that students of a second language have to master by the end of grade three. The first construct is oral language proficiency as it relates to English vocabulary development and the second relates to reading proficiency skills. The indices are constructed using principal component analysis (PCA), and then standardised on the control group mean and standard deviation. The English oral language proficiency index is constructed using the English expressive vocabulary task and the English listening comprehension task. The English reading proficiency index is constructed using the English word recognition, English oral reading fluency, English reading comprehension and English written comprehension subtasks.

The teacher questionnaires included questions on implementation fidelity from the teachers' perspective such as whether they attended ESL training, whether they received coaching support, the ESL materials that they received and the amount of time they spent a week on teaching ESL and HL. To evaluate instructional practice change we also asked teachers questions on the weekly frequency with which they implement certain activities and the resources they use during their lessons. Fieldworkers were also required to rate availability and quality of reading resources in the classroom, such as posters, flashcards, and reading books. We combined these indicators to construct a Kling index for classroom quality.

Three additional evaluation activities were conducted at the end of the third year of implementation, each aimed at providing a different perspective on the mechanisms which contributed to the success of the interventions. The first activity entailed retesting a sub-sample of the students who were assessed in the main data collection activity, as a fieldworker quality check. For these students we administered an ex-

tended vocabulary assessment and re-tested the students on five of the sub-tasks in the main assessment. The re-test and extended vocabulary assessments were administered by a different set of fieldworkers on six students per school from the main sample, and were conducted on the same day as the main data collection. The sample of students was pre-selected by the evaluation team and included two students at the top, middle and bottom of the performance distribution. The purpose of the re-test 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 student vocabulary development in both HL and ESL. 315 students from 60 schools participated in the vocabulary and re-test assessment. Comparison between the main data collection and re-test data gives us confidence that the inter-rater reliability is high. Table A.4 shows that the difference in the mean value between the original and re-test data for each subtask is statistically indistinguishable from zero, and that the correlation coefficients are high, ranging between 0.80 and 0.92.

The second activity was a classroom observation study that had well-trained fieldworkers (all currently pursuing a post-graduate degree) observe both the HL and ESL lessons of 53 schools in the sample, during the third term of the third year of the study. We randomly sampled 20 teachers in each treatment arm —stratifying by the language of instruction in the school (isiZulu or Siswati) and baseline learning outcomes. Due to protest action that was unrelated to the research study, we were unable to observe the lessons in two control schools, three on-site schools and two virtual schools. The classroom observation instrument was specifically designed for the purpose of the study, and the fieldworkers recorded how teachers were performing the different learning exercises required by the curriculum: vocabulary development, phonics and phonemic awareness, shared reading, group-guided reading and writing. The fieldworker also took a snapshot of teaching behavior at two different points in the lesson —at minutes 15 and 40 of the lesson — and recorded if the teacher was doing any of the following: giving instructions, listening to students read, reading to students, writing on the board, working with individual students, handing out books, doing admin at her desk or other non-teaching activities. Fieldworkers also observed the HL lesson in the same school. Since not all teachers teach both HL and ESL, this sample is further restricted to 44 teachers who teaches both (13, 15 and 16 teachers in the control, on-site and virtual arms respectively).

In addition to the lessons observed, the researchers also conducted a more in-depth

document review of students' written exercises, as well as interviews with the teachers. These interviews allowed us to ask more in-depth questions about the intervention, coded by high-quality enumerators. Importantly, the enumerators were trained to record if the teachers brought up the EGRS intervention when asked open-ended questions such as: (i) what has helped you most in covering the curriculum this year; and (ii) who checks that you are completing the curriculum?

Finally, for the virtual arm we also have access to rich tablet usage data, which has records of every occasion teachers accessed any particular slide or watched a video on the tablet. Due to some challenges in extracting this data, the most complete dataset exists for term 3 of 2019. This was the third year of the intervention, in which grade three teachers were receiving support. Figure A.6 provides schematic summary of the timeline of the intervention and data collection, and Figure A.5 provides a summary of sample selection, take-up, and attrition across all evaluation arms and waves of data collection.

4.2 Descriptive statistics, balance and attrition

Tables A.5 to A.7 provide some basic descriptive statistics of the sample, and show that the sample is balanced on a range of school, teacher and student characteristics, respectively. As to be expected, the majority of the schools are rural (74.4 percent) and fall in the lowest official school poverty quintile (53.9 percent). The teachers are relatively well-educated— 70 percent have at least a bachelors degree— and are mostly female. The average class size is quite large: 43 students per class. 29 percent of students are in a school where the language of instruction is isiZulu, whereas the other 71 percent are in Siswati schools. Table A.8 shows that the sub-sample of 53 school where we were able to conduct classroom observations is also balanced on the same set of characteristics. Figure A.7 shows kernel density plots of ESL Oral and Reading Proficiency, as well as HL reading proficiency.[13] It is encouraging that there are no large floor or ceiling effects, implying that our outcome measures discern proficiency across the full distribution of student ability.

Table A.10, column (1), shows that the attrition rate is 18 percent in the control, and balanced across treatment arms. Moreover, columns (2) to (5) show that treat-

[13]See Table A.9 for the descriptive statistics of each assessment instrument administered to students at baseline.

ments do not change the composition of attriters, relative to the control. Table A.11 shows that the sample remains balanced if we exclude the attriters. It is therefore unlikely that attrition would bias the results. Table A.12, column (1) shows that 68 percent of the original sample of grade one students (and 83% of the non-attriters) reached grade three by the third year of the study. Surprisingly, students in both treatment arms were 5 percentage points *less* likely to reach grade three.[14] Columns (2) to (5) show that older students, girls, and those who scored higher on the baseline assessment were more likely to have reached grade three.

4.3 Empirical strategy

Our main estimating equation is:

$$y_{icsb1} = \beta_0 + \beta_1(\text{On-site})_s + \beta_2(\text{Virtual})_s + X'_{isb0}\Gamma + \rho_b + \varepsilon_{icsb1}, \quad (1)$$

where y_{icsb1} is the endline (end of third year) outcome variable for student i who is taught by a teacher in class c, school s and strata b; $(\text{On-site})_s$ and $(\text{Virtual})_s$ are dummy variables indicating treatment status; ρ_b refers to strata fixed effects; X_{icsb0} is a vector of baseline controls; and ε_{icsb1} is the error term clustered at the school level. The controls include: the students' scores on the baseline sub-tasks, student gender, student age, the education district, the quintile of the socio-economic status of the school, and fieldworker fixed effects.[15] Moreover, since attrition was not uniform across schools, we also re-weight each observation based on number of students so that each school has an equal weight in the regressions. Results are robust to the exclusion of these weights. Some analysis is also at the teacher and school level. For these specifications we only include the strata as controls.

[14] One possible reason for the lower grade promotion in the treatment arms is that students at the bottom of the distribution learnt less as a result of the program. We investigate this further in section 5.2.

[15] We selected these controls prior to estimating the treatment effect on the full sample. We did this by restricting ourselves to the control data, and regressing the main outcome on the control variables, iteratively adding more controls. We only chose controls that substantially increased the R^2.

5 Results

5.1 Quality of Implementation

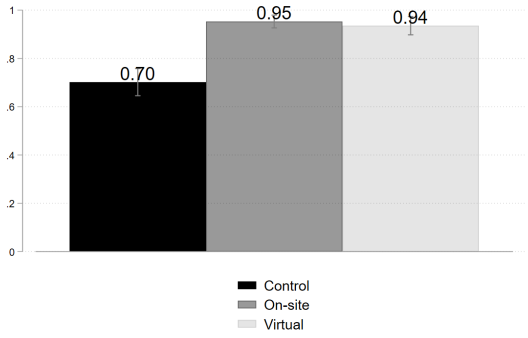
We start our presentation of results by examining the quality of implementation, which was high for both interventions for all three years of the program. Figure 1 shows high levels of teachers' exposure to key components common to both programs: attending training, receipt of lesson plans (either tablet or paper-based), and receipt of graded reading booklets. Nearly all teachers reported that they attended training in ESL (95 and 94 percent in the on-site and virtual arms respectively respectively), received graded reading booklets (96 and 95 percent), and are *using* the graded reading booklets (93 and 94 percent). A high proportion also reported using lesson plans provided by the government or a non-government organisation (85 and 82 percent respectively). Table A.13 reports these outcomes broken down by year of data collection and teacher grade level, which shows that the quality of implementation was consistently high over the full duration of the program. Table A.13 further shows that an index of classroom quality was substantially higher in both the on-site and virtual arms, in all three years of the study, reflecting the fact that teachers were displaying and using the additional learning aids provided by the program.

It is important to note that professional development activities were also taking place in the control schools. 70 percent of teachers in the control group report to have received training for ESL the same year they were exposed to the program, which was most likely provided by the province or the district. 49 percent of control teachers also reported to use ESL lesson plans that were provided to them by government or a non-government organisation. Although it is difficult to know the quality and type of training that was typically received in the control group, it is important to note that the counterfactual for this evaluation is schools and teachers that already receive some level of professional development support.

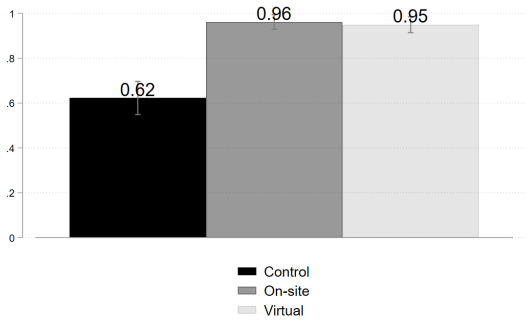
5.2 Learning

Next, we examine the impact on our primary and secondary outcomes of interest: English reading proficiency and oral language proficiency, respectively. Table 1, columns (1) and (6), show that by the end of the third year the on-site coaching program improved students' English reading and oral language proficiency by 0.13 and 0.31

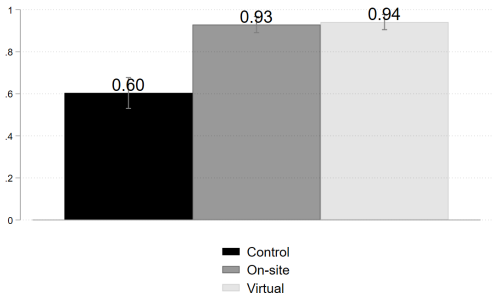
Figure 1: Quality of Implementation



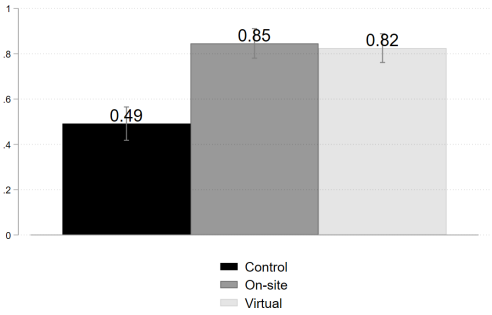
(a) Received training



(b) Access to graded reading booklets



(c) Use graded reading booklets



(d) Use lesson plan

Note. Data from teacher questionnaires. 307 grade one teachers were surveyed the first year of the study, 301 grade two teachers the second year, and 296 grade three teachers the third year. Panel (a) uses data from all grades. Panels (b) to (d) include data for the grade 2 and grade 3 teachers only. Moving from left to right, the bars indicate the average in the Control, on-site, and virtual arms respectively. Lines show 95 percent confidence intervals, with standard errors clustered at the school level.

standard deviations, respectively. These results are statistically significant at the 10 percent and the one percent levels. In contrast, the virtual coaching group only improved English oral language proficiency by 0.12 standard deviations —less than half the magnitude, relative to the on-site coach— and had no statistically detectable impact on reading proficiency skills. Moreover, the difference in effect sizes between the on-site coaching arm and the virtual coaching arm is statistically significant at a 5 percent level, for both outcomes.[16] Section A.1 in the appendix presents a series of robustness checks, which demonstrate that the results are not driven by differential or non-random attrition (Tables A.14 and A.15), are robust to the exclusion of student-level weights (Table A.16), and do not depend on the choice of student assessments instrument (Table A.17).

The remaining columns in Table 1 show results for each sub-task that constitute the two indices. Students in the on-site arm can read 2.66 more words on average relative to the control —a 12 percent increase— and their performance in the comprehension test improved by 6 percentage points— a 31 percent increase. There is no statistically significant impact on oral reading fluency or the written comprehension test. It is encouraging that there is a statistically significant impact on both listening and reading comprehension, since these are arguably the most important outcome indicators for a second language learner. In contrast, students in the virtual arm do not perform better in any sub-task related to reading proficiency, relative to the control, and effect sizes on vocabulary and oral comprehension are small: less than half the size of the on-site arm.

One way to interpret the magnitude of the impacts is to compare it to gains in the control over the period of the treatment. Although we did not assess English reading comprehension at baseline, we can place a lower bound on the learning if we conservatively assume that the entire stock of achievement developed over the three years of school. With this assumption, the improvements in English comprehension are at least 31 percent of the cumulative learning in the control over the three years of the intervention.[17] Nonetheless, performance in the on-site arm remains weak, with an average score of 25 percent in the comprehension test.

Although it is encouraging that the virtual arm had an impact on oral language

[16]The sharpened q-values for the three hypotheses tested on our primary outcome in column (1), using the procedure for controlling for the False Discovery Rate proposed by Anderson (2008), are: 0.063 ($\hat{\beta}_1 = 0$), 0.497 ($\hat{\beta}_2 = 0$), and 0.059 (($\hat{\beta}_1 = \hat{\beta}_2$)

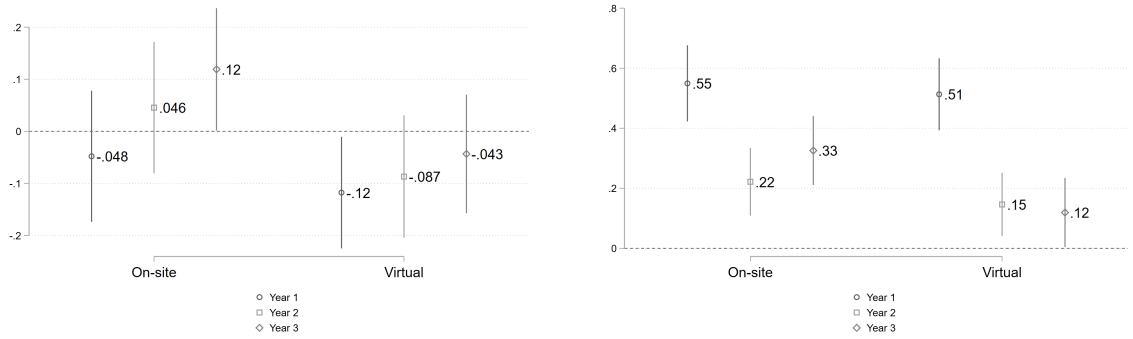
[17]$0.06/0.19 = 0.31$

Table 1: Impacts on student learning

	Reading proficiency					Oral proficiency		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Index	Word recog.	Read. fluency	Read. compr.	Written compr.	Index	Vocab	Listen compr.
On-site coach	0.130	2.660	2.458	0.060	0.016	0.313	0.420	0.074
	(0.068)	(1.360)	(1.909)	(0.019)	(0.020)	(0.068)	(0.105)	(0.017)
Virtual coach	-0.047	-1.199	-0.818	0.016	-0.019	0.123	0.147	0.032
	(0.069)	(1.357)	(1.902)	(0.018)	(0.019)	(0.072)	(0.118)	(0.018)
Control mean	0.000	23.121	27.255	0.191	0.355	0.000	3.120	0.216
Observations	2632	2684	2684	2684	2632	2684	2684	2684
R-squared	0.299	0.254	0.265	0.264	0.218	0.295	0.266	0.231
F-Test	0.019	0.009	0.108	0.037	0.112	0.020	0.032	0.036

Notes: Each column is a separate regression, estimated using equation 1. Standard errors are clustered at the school level. Estimates include strata and enumerator fixed effects and the following controls: students' scores on the baseline sub-tasks; gender and age; district; school's socio-economic status. The final row reports the p-value of the F-test of equality of coefficients. Indices are constructed using principal component analysis, and standardized to have control mean of zero and standard deviation of one. See table A.3 for the subtasks that go into each index. Vocabulary and listening comprehension are measured by the number correct; word recognition and oral language fluency are measured in words correct per minute. Listening, reading and written comprehension are measured as proportion of questions correctly answered.

Figure 2: Evolution of effect sizes over time



(a) English Reading Proficiency (b) English Oral Proficiency

Note. Coefficient plots, estimated using equation 1. A different regression is run for each outcome and for each year of data collection. See Table 1 for choice of controls, and Table A.3 for the subtasks that go into each index for each year of data collection. Data is restricted to the 2,403 students who were assessed in every round of data collection. Lines indicate 90 percent confidence intervals.

proficiency, these gains are associated with activities which are introduced earlier in the curriculum. The focus of the grade one curriculum is on oral language proficiency, with reading proficiency being introduced in grade two and receiving a stronger focus in grade three. Indeed, Figure 2(b) shows that both programs had large impacts of similar magnitude on oral language proficiency—0.55 and 0.51 standard deviations respectively— in the first year of the intervention, but the magnitudes decreased over time as the teaching shifted towards teaching decoding skills.[18] Similarly, Figure 2(a) show that neither program had a positive impact on reading proficiency at the first year, but in the on-site arm there was a gradual increase in effect sizes over time. This suggests that the virtual coach was successful at facilitating teachers' adoption of teaching practices aimed to vocabulary development, but not decoding. We discuss this in more detail in section 5.3 below. Note that since a different cohort of teachers were exposed each year, and data collection took place at roughly the same period each year, these dynamic impacts are unlikely to be due to dynamic responses in teachers' effort levels and/or learning over time.

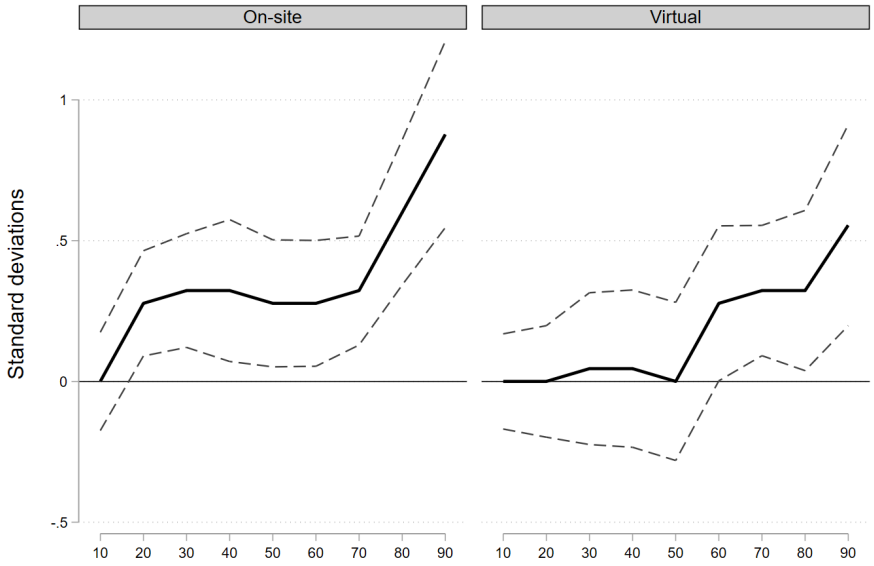
[18]The mean indices are not directly comparable across years, since the learning assessments were calibrated each year to discern across the distribution of student ability. The reductions in magnitudes in terms of standard deviations could be explained by the fact that control schools are catching up to the treatment schools. Also, see Tables A.18 and A.19 for comparison of effect sizes across time for the substasks that were administered both in the third year and in previous years.

Next, we explore the distribution of effect sizes, and find that the best-performing students consistently benefited most. Figures 3 and 4 report quantile treatment effects for each decile of performance in English oral and reading proficiency respectively. There is a consistent pattern for both programs and both outcomes that students at the top of the distribution —especially students in the top two deciles— benefited more from the programs, relative to students further down in the distribution. In fact, Figure 3 suggests that only the upper half of students in the virtual arm improved their oral proficiency as a result of the program, relative to equally-ranked students in the control. Similarly, Figure 4 suggests that only the upper half of students in the on-site arm improved their reading proficiency. More worrisome, the reading proficiency for students in the bottom 40 percent of the distribution actually *reduced* as a result of the virtual coaching program.[19] In other words, only the best-performing students benefited from the on-site coaching program, and the worst-performing students suffered from the virtual coach. This suggests that the program might still be targeted at a level higher than the median student, reflecting the possibility that the curriculum in South Africa assumes a higher proficiency amongst learners entering each grade than is currently the reality. Requiring students to adhere to that curriculum, without sufficient remediation from teachers, could disadvantage the weakest students. Section A.2 (Table A.20 and Figures A.10 and A.11) in the appendix shows that these results are broadly confirmed in a regression framework, interacting treatment status with the index for baseline reading proficiency.

Moving beyond English literacy, we also assessed students' home language literacy and mathematics skills to evaluate whether the treatments had any crowding-out or spillover effects on the other subject areas. Table 2 shows a *negative* estimated effect of the virtual coaching program on home language literacy of 0.19 standard deviations by the end of the third year of the study. There is also a significant reduction in home language oral reading proficiency, reading comprehension and written comprehension. In contrast, there is no negative impact of the on-site arm on the reading index, although there is a statistically significant negative impact on home language oral reading proficiency and reading comprehension. Moreover, the negative effects across the sub-tasks are consistently larger for the virtual treatment arm relative to the

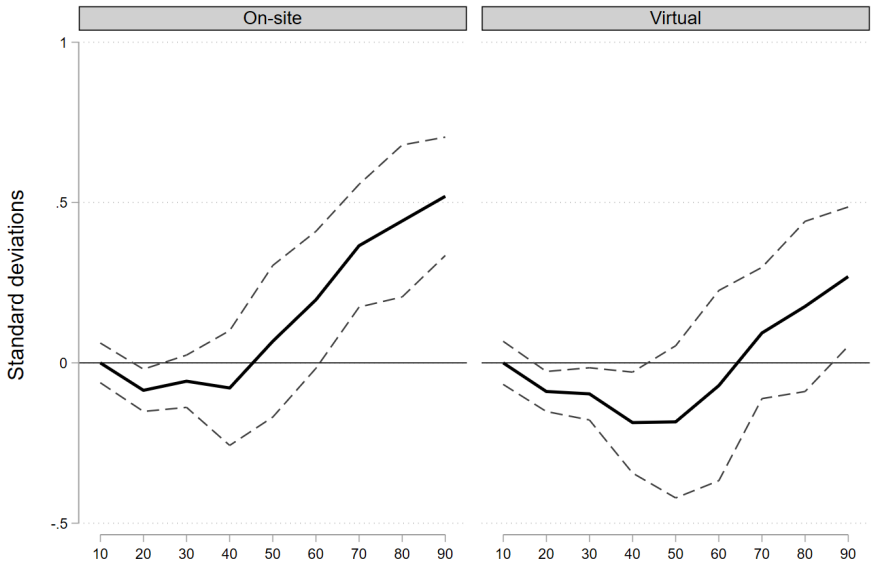
[19]The U-shape of the quantile regressions for reading proficiency is due to floor effects: 9, 10, and 12 percent of students in the control, on-site and virtual arms respectively could not read a single word in English. See Figure A.9 for a comparison of cumulative density functions of reading proficiency by treatment arm

Figure 3: Quantile Regressions— English Oral Proficiency



Note. Coefficient estimates of unconditional quantile regressions for each decile of student performance, with standard errors clustered at the school level. Confidence intervals are 90%. The bottom decile is on the left-hand side, and the top decile is on the right-hand side.

Figure 4: Quantile Regressions— English Reading Proficiency



Note. See Figure 3.

Table 2: Home Language Literacy and Numeracy

	Home Language					Maths
	(1)	(2)	(3)	(4)	(5)	(6)
	Index	Letter recog.	Reading fluency	Reading compr.	Written compr.	Maths
On-site coach	-0.047	4.850	-2.393	-0.032	-0.035	0.016
	(0.068)	(1.885)	(1.155)	(0.022)	(0.019)	(0.020)
Virtual coach	-0.193	-0.973	-3.021	-0.066	-0.068	-0.019
	(0.074)	(1.772)	(1.285)	(0.023)	(0.020)	(0.019)
Control mean	0.000	42.947	23.091	0.480	0.407	0.355
Observations	2632	2684	2684	2684	2632	2632
R-squared	0.290	0.238	0.245	0.255	0.240	0.218
F-test	0.059	0.004	0.625	0.183	0.101	0.112

Notes. See table 1. Letter recognition, word recognition, and oral reading fluency are measured as the number correct per minute; reading comprehension, written comprehension, and mathematics are measured as proportion of questions correctly answered.

on-site arm, and the difference in the mean index is statistically significant at the 10 percent level. There is no impact, either positive or negative, on mathematics. We discuss possible reasons for this result in section 6.1 below.

5.3 Teaching practice

Finally, we investigate whether teaching practices changed as a result of the interventions. Table 3 shows that there was a large shift in observed teacher instructional practices, for both programs. According to the classroom observations, teachers in both treatment arms were *less* likely to teach vocabulary or phonics, but *more* likely to have the students practice writing, relative to the control. This is in line with curriculum expectations. Moreover, teachers in the on-site arm were more likely to practice group-guided reading, relative to the control, and their students were almost fives times more likely to get a chance to read out loud individually during the lesson. This was not the case for teachers in the virtual arm. Section A.3 in the Appendix shows that these results are broadly mirrored in the teacher survey data (Table A.21), and also provides suggestive evidence of positive spillovers of teaching practices into teaching of home language literacy (Table A.22).

Table 3: Learning activities observed during the English lesson

	(1) Language Phonics	(2) Shared reading	(3) Group-guided reading	(4) Pupil reads individually	(5) Writing
On-site	-0.184	-0.023	0.293	0.333	0.276
	(0.103)	(0.142)	(0.148)	(0.121)	(0.115)
Virtual	-0.179	-0.296	0.175	0.133	0.238
	(0.099)	(0.149)	(0.156)	(0.120)	(0.124)
Observations	53	53	53	53	53
R-squared	0.208	0.309	0.268	0.345	0.230
Control mean	1.000	0.778	0.167	0.056	0.722
F-test	0.975	0.104	0.442	0.157	0.492

Notes. Each column is a separate regression. Data comes from classroom observations conducted when the teacher was teaching ESL. Outcomes are dummy variables equal to one if the respective teaching activities took place at least once during the full duration of the lesson. Estimates include strata fixed effects, and standard errors are clustered at the school level. The final row reports the p-value of the F-test of equality of coefficients

The fact that the on-site, but not the virtual, coaching program had a large positive impact on group-guided reading could explain why both programs improved oral language proficiency, but only the on-site coaching program improved reading proficiency. Group-guided reading is an important input into learning, since gives the teacher an opportunity to provide more individualized feedback to each student, but is a difficult technique to implement.[20] To test the hypothesis that group-guided reading is an important input for acquisition of reading proficiency, we conduct mediation analysis, using the sequential g-estimation as developed by Acharya et al. (2016) (see section A.4 in the Appendix). This approach requires strong assumptions for identification. Nonetheless, Table A.23 provides suggestive evidence that as much as 59 percent of the improvements in reading proficiency in the on-site arm could be explained by the increased usage of group-guided reading.

[20]It requires more complex interactions with students as well as good classroom management to ensure that the students who are not in the small group are being quiet and productive. In fact, 46 percent of teachers in the control consider group-guided reading to be hard, compared to only 19 and 23 percent respectively for teaching phonics and conducting shared reading.

6 Discussion

In this section we explore possible reasons for the unexpected negative effects on home language literacy, we investigate why the on-site coaching program was more effective than the virtual coaching program, and we perform a cost-benefit analysis. Since none of the evidence reported in this section was specified in our pre-analysis plan (with the exception of Figure 6), this analysis should be considered exploratory.

6.1 Why was there a negative impact by the virtual arm on home language?

A priori, the direction of the impact of the programs on home language could be either positive or negative. On the one hand, there could be a negative impact if there is a crowding out of teaching time. This would be the case if the lesson plans require additional work, but the teacher is not able to complete all the content in the lesson plans within the allocated time. Moreover, there could be a crowding out of teacher professional development in other subjects: teachers in the intervention schools are spending all of their professional development time on this program, so might be receiving less training in other foundation phase subjects, relative to the control. On the other hand, a positive impact on learning in other subjects is also possible if the improved teaching practices adopted by teachers during the ESL classes are applied to the teaching of other subjects. Moreover, students' home language reading proficiency could also improve, if there is a transference of phonemic awareness and decoding skills between the two languages, provided that both the teacher and students have sufficient knowledge of the orthographic rules for both languages.[21]

It is unlikely that the crowding out of home language professional development explains the result. Table 4 shows that grade three teachers in the virtual arm were not significantly less likely to receive training in home language in 2019 (the year of the intervention), nor is there any difference in the proportion of teachers who have graded reading booklets or lesson plans for home language instruction. Moreover, if the control teachers benefited more from professional development in home language instruction, one would expect to also observe improved pedagogical practices in the

[21]Note that the data do not allow us to conclusively rule out a mechanism of transference of reading skills across languages. Future research will examine this question in more detail, drawing from evidence across a range of different studies.

control relative to the intervention teachers. But results from classroom observations of teaching of home language suggests the opposite: Table A.22 shows that grade 3 teachers' teaching practices in home language are slightly better in the on-site arm relative to the control, and no different in the virtual arm.

Columns (4) and (5) in Table 4 provides some evidence of a crowding out of teaching time, especially in the virtual arm. As mentioned in section 2.2, the official curriculum allows teachers to decide between teaching three or four hours of English a week (which will result in eight hours or seven hours of home language, respectively). The lesson plans used in this study, however, specified that teachers had to spend four hours teaching English and thus only seven hours teaching Home Language Literacy. It is thus possible that the programs (intentionally) caused a shift in teaching time away from Home Language to English. Column (4) in Table 4 shows that teachers in both interventions reported spending less time a week teaching home language, but the magnitude of the reduction is small: teachers in the virtual arm reported dedicating on average 18 fewer minutes to home language instruction per week. Note that teachers in the control schools *already* dedicated just under the minimum requirement of seven hours to home language. This suggests that any observed reduction goes beyond what is intended by the interventions. Indeed, column (5) shows that the teachers in the virtual arm in particular are almost twice as likely to report to spend less than the minimum requirement of seven hours of teaching home language. As a result, 41 percent of teachers in the virtual arm allocate fewer than seven hours per week to home language instruction. There is no statistically significant increase in this probability the on-site arm.

Results from the survey administered during the classroom observations provide additional insights into why teachers in the virtual arm dedicate less time to home language. Figure A.12 shows that teachers in the virtual arm were less likely to be satisfied with how much they have progressed in the home language curriculum in the year of the study (71 percent versus 93 percent in the control), and all of these teachers refered to this program when explaining why they are struggling to complete the curriculum.[22] Given the small sample of teachers surveyed in the classroom observations who teach both ESL and HL the difference is not statistically significant,

[22]Examples include: They do not get the same kind of support as they get for teaching ESL, teaching ESL takes time away from teaching in the home language, and the teacher finds teaching home language more challenging because it is not on the tablet.

Table 4: Investigating spillovers

	HL Professional Development			HL Instruction Time	
	(1)	(2)	(3)	(4)	(5)
	Training	Lesson plans	Graded readers	Total hours	< 7 hours
On-site	-0.157	0.115	0.010	-0.216	0.123
	(0.075)	(0.070)	(0.073)	(0.120)	(0.076)
Virtual	-0.074	0.007	-0.024	-0.301	0.215
	(0.078)	(0.061)	(0.076)	(0.116)	(0.074)
Control mean	0.526	0.183	0.637	6.980	0.228
Observations	292	281	278	281	281
R-squared	0.098	0.041	0.091	0.073	0.109
F-test	0.340	0.159	0.674	0.478	0.287

Notes. Each column is a separate regression. Data come from grade three teacher survey. The outcome variables in the first three columns are dummy variables equal to one if a teacher (i) received professional development support in home language in 2019, (ii) uses HL lesson plans provided by an NGO, and (iii) has HL graded reading booklets, respectively. The outcome variable in the fourth column is the total number of hours that teachers report to allocate to HL instruction in a week. In the fifth column it is a binary variable equal to one if a teacher reported to allocate fewer than seven hours a week to HL instruction. Estimates include strata fixed effects, and standard errors are clustered at the school level. The final row reports the p-value of the F-test of equality of coefficients

making this merely suggestive evidence.

In sum, we find evidence that crowding out of teaching time in the virtual arm forms at least part of the explanation for students' weaker performance in home language literacy. It is possible that the teachers found it challenging to complete all the activities required by the lesson plans. Given that the lesson plans were closely aligned to the official curriculum, the implication is that adhering closely to the activities required by the curriculum within the allocated time may be a challenge. Why then do we not observe the same degree of crowding out of teaching time or a similar negative impact on HL learning in the on-site coaching group? One possibility is that the targeted nature of support possible through in-person visits helped with time management and helped mitigate against borrowing time from HL lessons. We turn to this below.

6.2 Why was the virtual coaching intervention less effective?

Although our study design does not allow us to conclusively prove that the three mechanisms enabled through in-person interaction —i.e. targeted feedback, and development of relationships of accountability and trust— are the primary drivers for the differences in impacts, this section provides evidence that alternate explanations —such as differences in the quality of implementation, differences frequencies of interaction with the coach, and barriers to accessing the technology— are unlikely to explain the differences. We also present empirical evidence to the additional accountability and support provided by the in-person visits.

First, it is unlikely that differences in the quality of implementation explains the results. The same organization implemented both interventions, and we demonstrate in detail in section 5.1 that the quality of implementation for both programs was high. Moreover, both interventions were equally effective at improving oral language proficiency after the first year of the intervention.

Second, it seems equally unlikely that differences in the length of exposure to a coach explain the result. The average number of times that a teacher received a phone call by a coach in the virtual arm is 10, slightly fewer than the 14 times that a teacher was visited by a coach in the on-site arm. However teachers in the virtual arm also received weekly text message reminders from the coach, and teachers could also call or text the coach if they had specific questions, and had the option to watch instructional

videos. Our reading of the literature makes us believe that it is unlikely that such a small difference in the length of exposure to a coach can explain why the virtual coach had no impact on reading proficiency. For comparison, authors in the first early-grade reading study in South Africa found a large positive significant impact on learning for students of teachers were visited on average 10 times during the year, so it is unlikely that the impacts of on-site coach would be zero if the number of visits go down from 14 to 10 (Cilliers et al., 2019). Moreover, in a randomized evaluation, Piper and Zuilkowski (2015) found that there is no statistically significant difference in impacts on English language if a coach is responsible for serving 10 schools rather than 15, thus visiting teachers more frequently during the year. In a meta-analysis of coaching programs, Kraft et al. (2018) found no relationship between the effect size of a program and the total hours of exposure between the coach and the teacher.

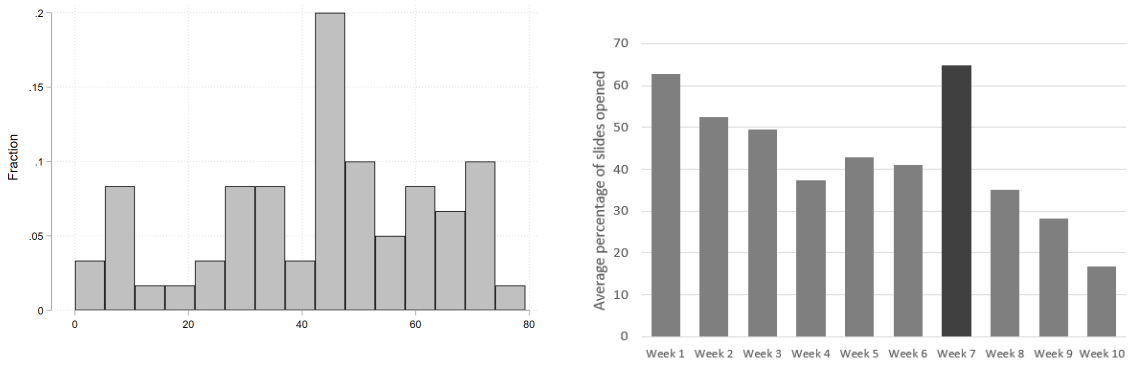
Third, analysis of tablet usage data suggests there were no barriers to accessing the technology: almost all teachers used the tablets, although at a variable rate. Panel (a) in Figure 5 shows a histogram of the distribution of percentage of term 3 lesson plan slides that were accessed by teachers any time between June and September 2019.[23] This might be considered a crude measure for potential curriculum coverage, or alternatively a proxy for intervention implementation fidelity. There is clearly a high variation in accessing slides, but notably only two teachers (3.3 percent) did not open a single slide during the third term. This is also consistent with our findings during the classroom observations interview, where 88 percent of teachers reported that they are very comfortable with it, and 12 percent of teachers reported that they are somewhat comfortable with the tablet.[24]

Rather, the pattern of slide coverage over the duration of the semester suggests that teacher motivation or ability to complete the curriculum was the binding constraint, and not the technology itself. Figure 5, panel (b), shows that there was a gradual decline in the proportion of slides covered in a week, with the lowest coverage seen in the final week and the higher coverage in the beginning of the term. The one exception to this trend was week seven, which saw the highest coverage rate. This

[23]The paper-based lesson plans were reformatted into pdf slides for teachers to navigate through on the tablet.

[24]The fact that 96.7 percent of teachers accessed the slides means that older teachers did not face barriers to opening the slides. As an additional test to rule out age as a constraint, Table A.24 shows that older teachers in the virtual arm are no less likely to use lesson plans relative to the control, even though these lesson plans are on the tablets.

Figure 5: Proportion of slides accessed on the tablet



(a) Histogram (b) By week

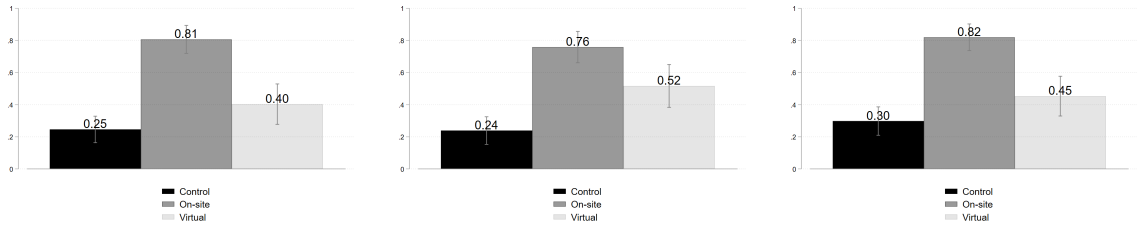
Note. Tablet usage data. The paper-based lesson plans were reformatted into pdf slides for teachers to navigate through on the tablet. Panel (a) shows a histogram of the proportion of slides that were opened by each teacher in the on-site arm, between July and September 2019. Panel (b) shows the proportion by each week over that same period.

also happens to be the week when assessments should take place.[25] The fact that teachers' usage of the lesson plans provided in the tablets almost doubled when they faced stronger incentives suggests that the teachers in the virtual arm are still far from their production possibility frontier.

Finally, results from the teacher questionnaire and interviews conducted after the classroom observations provide supporting evidence that the on-site coach played an important role in holding the teachers accountable, and teachers were more likely to turn to them for support. Figure 6 shows that teachers in the on-site arm were more likely than both the control teachers and the virtual arm teachers to respond that (i) they had been observed by a coach at least twice this year, that (ii) a coach modelled a lesson for them at least twice this year, and that (iii) they received a compliment from a coach. Teachers supported by the virtual coach were also more likely than the control teachers to have responded positively to these questions, but the magnitudes are substantially smaller. In addition, Figure 7 shows that teachers in the on-site arm were more likely to mention the coach as someone who checks if she is completing the curriculum, and more likely to mention the coach as someone who has helped her learn most this year. Consistent with this result, teachers in the on-site arm were

[25] Teachers are expected to upload assessment results onto SA-SAMS, a government wide school management system into which teachers have to upload various data.

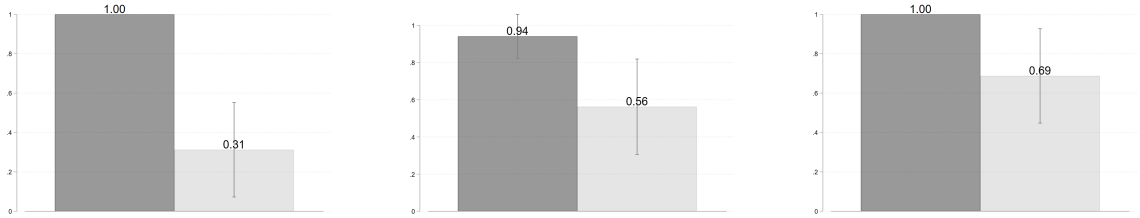
Figure 6: Support received by a mentor or coach



(a) Observed teaching (b) Model lesson (c) Received compliment

Note. Data from teacher questionnaires administered to 296 grade three teachers in the final year of the study.

Figure 7: Coach accountability and support in completing curriculum



(a) Coach checks that teacher is completing curriculum (b) Coach supports teacher in completing curriculum (c) Teacher is satisfied with curriculum coverage

Note. Data from the teacher interview held with 53 teachers in 53 schools after the completion of the classroom observations. From left to right, the bars indicate averages in the on-site and virtual arms respectively. Confidence intervals are at a 95 percent level

more satisfied with their curriculum coverage.

6.3 Accounting for Hawthorne effects

Since we have a panel of students who were tracked over a period of three years, there is a potential concern that teachers prioritize the learning of these students, if they come to learn that the same group of students are assessed every year. This could lead to an upward bias of our results, if teachers have incentives to impress the research organization by demonstrating higher levels of learning for these students, *and* if these incentives are stronger in the treatment arms. We believe that such a bias is unlikely given our context, for two reasons. First, a new group of teachers were exposed to the program each year, so it is highly unlikely that the new group of

teachers know which students were assessed the prior year. It is only in the first year where students were assessed at both the beginning and the end of the year, but the teachers were not informed that the same group of students will be assessed again at the end of the year. Second, there was a clear distinction between the program and the data collection. The program implementer and data collection companies were different organizations with their own unique brands. The head teacher knew that the purpose of the data collection was to assess the program, but there is no reason to believe that the teachers associated data collection with the programs.

6.4 Cost-effectiveness

Next, we compare the cost effectiveness of the two coaching modalities. For cost estimates, the expenditure data for the three years of implementation was taken, excluding any costs that were involved in the development and piloting of the program.[26] We also do not include the cost of purchasing the tablets, but include depreciation costs, estimated as the sum of year's digits method and assuming that the tablets last for seven years.[27] These estimates should therefore provide a realistic per-student cost if these models of delivery were scaled up. Based on these estimates, the per student costs of on-site coaching was USD66 per year and USD52 for virtual coaching. In terms of the cost of supporting a teacher per year, it is USD2,750 for on-site coaching and USD2,168 for virtual coaching. To place this in context, the average yearly teacher salary in South Africa was USD36,572 in 2019. This means that the cost of supporting a teacher for a year is eight and six percent of their yearly salary for the on-site and virtual coaching programs respectively.

Given the impacts of 0.31 on oral language proficiency and 0.13 on reading proficiency for on-site coaching at the end of the study, there was a 0.16 standard deviation increase in oral language proficiency for each USD100 spent and a 0.08 increase in reading proficiency. For virtual coaching, there was no significant impact on reading

[26] Ongoing costs such as material revision and the development of new audio and video clips were still included since these resources are developed in response to the teaching challenges experienced by teachers. All USD rates are calculated at a Rand:USD exchange rate of R14 per USD. Following Dhaliwal et al. (2013) we do not use the Purchasing Power Parity (PPP) adjusted exchange rate, since we are more concerned about its costs in a developing country context than what it would cost in the United States. For comparison, the PPP exchange rate in 2019 was 6.67, so the costs in USD (PPP) would be roughly twice the size.

[27] Given this method, the proportion of the purchase cost incurred over the first three years is: (7+6+5)/(7+6+5+4+3+2+1)=0.642.

Table 5: Cost-effectiveness of the on-site and virtual coaching interventions

	On-site	**Virtual**
Costs per learner per year (USD)	66	52
Costs per teacher per year (USD)	2,750	2,168
Effect size on oral language proficiency per USD100	0.16	0.08
Effect size on reading proficiency per USD100	0.07	-

Note. Initial development and piloting costs and costs of tablets not included, but material revision and development of new audio and depreciation of tablets included.

proficiency, but for oral language proficiency there was a 0.07 increase in oral language proficiency for each USD100 spent. On-site coaching, therefore, does not only have a larger impact on learning outcomes, but it is also twice as as cost-effective as virtual coaching.

The smaller than expected difference between the costs of the on-site coach and the virtual coach is due to two reasons. First, some of the largest cost drivers —such as training, program management, and teaching materials— are the same across the interventions (see Table A.25 for a more detailed breakdown of costs). In fact, 49 percent of the costs in the on-site arm are costs that are also incurred in the virtual arm as well. So, even though the salary and transport costs for the on-site coach are over three times larger than the salary and communication costs for the virtual coach ($33 vs $11 per student per year), these costs are only a fraction of the overall costs. Second, the virtual arm has two additional costs not incurred by the on-site program: the additional day of training, and the provision of tablets and hosting of software for the virtual arm. Tablets are often thought to be less expensive since they can be used for multiple years, but this is not the case in our study. There are other ongoing costs that needs to be taken into account, such as a technical assistant to support teachers who experience technical problems with the tablet or application and the hosting of the application that was developed.

7 Conclusion

This study compares the effectiveness of a structured pedagogy program that was implemented through two different delivery models: providing teachers with paper-based lesson plans and support from an on-site coach, or providing teachers with

lesson plans on a tablet and support from a virtual coach. After one year of exposure, the two programs were equally effective at improving students' English oral language proficiency. After three years of exposure, only the on-site coaching program succeeded in improved students English reading proficiency, by 0.13 standard deviations. The virtual coaching program had no impact on English reading proficiency, and also reduced home language reading proficiency, probably due to a crowding out of teaching time. Teachers in the coaching arm also experienced larger gains in productivity, especially for teaching practices that develop students' reading skills. We further show that the use of technology was not a barrier, and provide suggestive evidence that the virtual coach faced greater barriers to developing a trusting relationship, holding teachers accountable to completing the curriculum, and providing targeted feedback based on in-classroom observations.

The main finding of this paper is sobering: a virtual coaching alternative, which was somewhat less expensive and considerably less reliant on human resources, did not have the same desired effect, and actually reduced home language literacy. 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 developing countries. 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.

Two general policy recommendations are worth highlighting. First, our research design and detailed data collection allows us to develop hypotheses for the modality of virtual coaching support which *might* be effective. Most likely, a more effective coaching program should involve a combination of some initial face-to-face coaching to establish the relationship, followed up with virtual coaching to sustain the instructional practice change. Moreover, teachers need to share video recordings of their teaching to the coach, in order to receive targeted feedback. But these recommendations will be difficult to implement in resource-constrained settings in developing countries, so the cost advantage relative to on-site coaches would shrink. Moreover,

this raises a more fundamental problem of motivating teachers to engage with the technology and submit videos to a coach.

Second, this study demonstrates that strong complementarities exist between technological interventions and the incentives faced by those who are required to adopt the technology. Technology provides opportunities to improve teacher productivity, provided that the teachers face the appropriate incentives to apply these technologies. Although the virtual coach can provide the same technical input as an on-site coach, they cannot provide the same level of accountability, since they are not directly monitoring the teachers in the classroom.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen, "Explaining causal findings without bias: Detecting and assessing direct effects," *American Political Science Review*, 2016, *110* (3), 512–529.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484), 1481–1495.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying education: Evidence from two randomized experiments in India," *The Quarterly Journal of Economics*, 2007, *122* (3), 1235–1264.

Beg, Sabrin A, Adrienne M Lucas, Waqas Halim, and Umar Saif, "Beyond the basics: Improving post-primary content delivery through classroom technology," Technical Report, National Bureau of Economic Research 2019.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane, "Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa," *Journal of Economic Perspectives*, 2017, *31* (4), 185–204.

Bruns, Barbara, Leandro Costa, and Nina Cunha, *Through the looking glass: can classroom observation and coaching improve teacher performance in Brazil?*, The World Bank, 2017.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor, "How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching," *Journal of Human Resources*, 2019, pp. 0618–9538R1.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch, "Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education," *Education policy in developing countries*, 2013, pp. 285–338.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Tony Brady, Piotr Gawron et al., "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in The Gambia," *Journal of Development Economics*, 2020, *148*, 102539.

Evans, David K and Anna Popova, "What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews," 2016.

Howie, Sarah J, Celeste Combrinck, Karen Roux, Mishack Tshele, Gabriel Mokoena, Nelladee McLeod Palane et al., "PIRLS Literacy 2016: South African Highlights Report (Grade 4)," Technical Report, Centre for Evaluation and Assessment (CEA) 2017.

Kerwin, Jason T and Rebecca L Thornton, "Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures," *Review of Economics and Statistics*, 2020, pp. 1–45.

Kotze, Janeli, Brahm Fleisch, and Stephen Taylor, "Alternative forms of early grade instructional coaching: Emerging evidence from field experiments in South Africa," *International Journal of Educational Development*, 2019, *66*, 203–213.

Kraft, Matthew A, David Blazar, and Dylan Hogan, "The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence," *Review of educational research*, 2018, *88* (4), 547–588.

Lee, David S, "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102.

McEwan, Patrick J, "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments," *Review of Educational Research*, 2015, *85* (3), 353–394.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian, "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 2019, *109* (4), 1426–60.

Piper, Benjamin and Stephanie Simmons Zuilkowski, "Teacher coaching in Kenya: Examining instructional support in public and nonformal schools," *Teaching and Teacher Education*, 2015, *47*, 173–183.

_ , _ , and Abel Mugenda, "Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative," *International Journal of Educational Development*, 2014, *37*, 11–21.

_ , _ , Dunston Kwayumba, and Carmen Strigel, "Does technology improve reading outcomes? Comparing the effectiveness and cost-effectiveness of ICT interventions for early grade reading in Kenya," *International Journal of Educational Development*, 2016, *49*, 204–214.

Popova, Anna, David K Evans, and Violeta Arancibia, "Training teachers on the job: What works and how to measure it," *The World Bank Working Paper*, 2016.

Powell, Douglas R, Karen E Diamond, Margaret R Burchinal, and Matthew J Koehler, "Effects of an early literacy professional development intervention on head start teachers and children.," *Journal of educational psychology*, 2010, *102* (2), 299.

Snilstveit, Birte, Emma Gallagher, Daniel Phillips, Martina Vojtkova, John Eyers, Dafni Skaldiou, Jennifer Stevenson, Ami Bhavsar, and Philip Davies, "Education interventions for improving the access to, and quality of, education in low and middle income countries: A systematic review," Technical Report, The Campbell Collaboration 2014.

Taylor, Nick, "Equity, efficiency and the development of South African schools," in "International handbook of school effectiveness and improvement," Springer, 2007, pp. 523–540.

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

A Appendix

A.1 Robustness checks

We perform four robustness checks, focused on the two outcomes measured at the end of the third year: English oral language and reading proficiency. First, Table A.14 demonstrates robustness to non-random attrition, by weighing observations by the inverse of the predicted probability of attriting (effectively placing a higher weight on observations that have similar observable characteristics as those who attrited). Including the weights makes almost no change to the coefficient estimates for the on-site arm —an increase of 0.007 standard deviations for English language, and no change for English literacy— and slightly reduces the coefficient estimate for the virtual arm on English language proficiency— by 0.025 standard deviations. Second, we also test for robustness to non-random attrition by constructing Lee (2009) bounds— trimming the top and bottom distributions of student learning for the control group by the difference in attrition rates as a proportion of the remaining sample in the control.[28] Table A.15 shows that the lower bounds for the treatment effects of the on-site arm on oral and reading proficiency are only 0.024 and 0.016 SD smaller respectively, and remain statistically significant at a 10 percent level. The lower bound for the estimated treatment effect of the virtual coaching program on oral reading proficiency decreases by 0.024, and is no longer statistically significant at conventional levels of significance.

Third, we show in Table A.16 that results are similar in magnitude and remain statistically significant when we do not weigh each regression by the inverse of the number of students assessed in each school at the end of year three. Fourth, Table A.17 shows that the treatment effects on English vocabulary do not depend on the choice of words used in the original instrument. We retested a subset of our sample using a more expanded set of words in the vocabulary test. It is encouraging that the estimated treatment effect is of similar magnitude (it is, in fact, slightly larger for both treatment arms) when using the expanded instrument.

[28] In particular, given the attrition rates of 18.2 and 20.2 percent in the control and treatment arms respectively, we need to trim $(20.2 - 18.23)/(1 - 18.23) = 2.39$ percent of the remaining control observations at endline. Since there is a small ceiling effect in the control, where more than 2.39 percent of students have the same bottom score, we randomly selected $1185 * 0.0239 = 28$ of the worse-performers to be trimmed.

A.2 Heterogeneous treatment effects by baseline reading proficiency.

Table A.20 shows results from a regression that interacts treatment status with the index for baseline reading proficiency. For both programs, students who performed better on the baseline learning assessment improved their reading skills by more, relative to students who performed worse at baseline. Figures A.10 and A.11 plot local polynomial regression estimates of the relationship between treatment effect sizes and a student's percentile rank in terms of baseline academic performance. For reading proficiency, the relationship is strikingly linear and upward sloping, especially in the case of the virtual arm. In fact, figure A.11(b) suggests that the bottom fifth of students in the virtual arm in terms of baseline learning might have learnt *less* as a result of the program.

A.3 Additional teacher-level outcomes

Table A.21 reports results on teaching practice, using data from the grade three teacher survey. It shows that teachers in the treatment arms were more likely to implement the teaching activities at the required weekly frequency. Since teachers are required to implement group-guided reading daily, but phonics three times a day, this means that teachers in the treatment arms were more likely to conduct group-guided reading, but *less* likely to teach phonics. In fact, teachers in the control group were more than twice as likely to state that they teach phonics daily.

There is suggestive evidence that teachers in both arms applied some of their improved teaching practices to home language instruction as well as the English lesson. Table A.22 shows that during the home language lesson, teachers in the on-site arm were 29.4 percentage points (33 percent) less likely to teach phonics relative to the control, they were 16.9 percentage points (71 percent) more likely to practice group-guided reading, and as a result students were 30 percentage points (391 percent) more likely to have been observed reading individually to a teacher. Teachers were also 18.2 percentage points (87 percent) more likely to be observed working individually with a student. These effect sizes are similar in magnitude to the findings from the classroom observations for ESL, but they are less precisely estimated due to the smaller sample. Note that there is no evidence that teaching practices were worse relative to the control.

A.4 Mediation analysis

We also conduct mediation analysis, using the sequential g-estimation as developed by Acharya et al. (2016). The demediated outcome variables are constructed in two steps. First, we estimate:[29]

$$y_{icsb1} = \beta_0 + \beta_1(\text{On-site})_s + \beta_2(\text{Virtual})_s + X'_{isb0}\Gamma + \rho_b + \beta_3 M_{cs} + \mathbf{Z_{cs}}'\Delta + \varepsilon_{icsb1} \quad (2)$$

where M_{cs} is the mediating variable of interest, and $\mathbf{Z_{cs}}$ is a vector of post-treatment potential confounders. In our case M_{cs} is a binary variable equal to one if a teacher reports to practice group-guided reading on a daily basis, standardized to have a control mean of zero. The vector of additional confounding variables are: (i) binary variables equal to one if a teacher reports to perform shared reading, creative writing, or phonics at the correct weekly frequency; (ii) a Kling index of the classroom quality, including the number of books, quality of flashcards, and quality of posters; and (iii) hours a week spent teaching home language and English respectively.

Next, we construct the demediated outcome variable, $\tilde{y_{icsb1}} = y_{icsb1} - \hat{\beta}_3 M_{cs}$. We then estimate the treatment effect on the demediated outcome, using equation 1. The treatment impacts on $\tilde{y_{icsb1}}$ can be interpreted as the Average Controlled Direct Effect(ACDE)— what the treatment impact *would* have been, if the value of the mediating variable was set to zero (in our case, this is the same as setting the mediating variable equal to the mean in the control). The difference in magnitudes between the treatment effects on y_{icsb1} and $\tilde{y_{icsb1}}$ can there therefore be interpreted as the indirect impact of the treatment through the mediator— i.e. the contribution of the mediator to the overall treatment impact.

Columns (1) and (3) in table A.23 show the mean treatment effects on our two outcomes of interest, restricted to the observations that we could match student with teacher-level data. Columns (2) and (4) show the treatment effects on the demediated outcomes. Comparing columns (1) and (2) shows that the treatment effect of the on-site coaching program on oral proficiency would only have been 23 percent lower, if the program had no impact on group-guided reading. The magnitude of the coefficient in column (2) remains large. In contrast, a comparison of columns (3) and (4) reveal that the treatment effect on reading proficiency would have been 59 percent lower if there

[29] This model is equivalent to equation 1, except for the addition of two terms, M_{cs} and $\mathbf{Z_{cs}}$

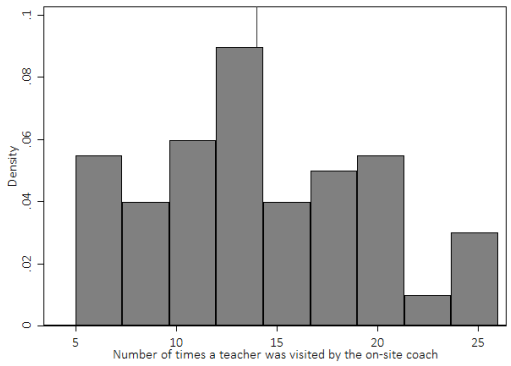
were no impact on group-guided reading, and no longer statistically significant.[30]

[30](0.175-.072)/0.175=0.59

A.5 Supplementary figures and tables

Figure A.1: Histogram of number of times that a teacher was visited or called by a coach

(a) on-site



(b) virtual

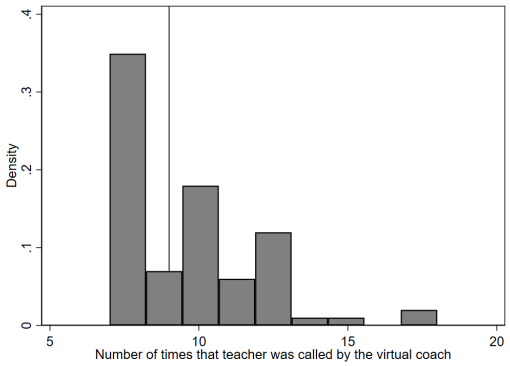
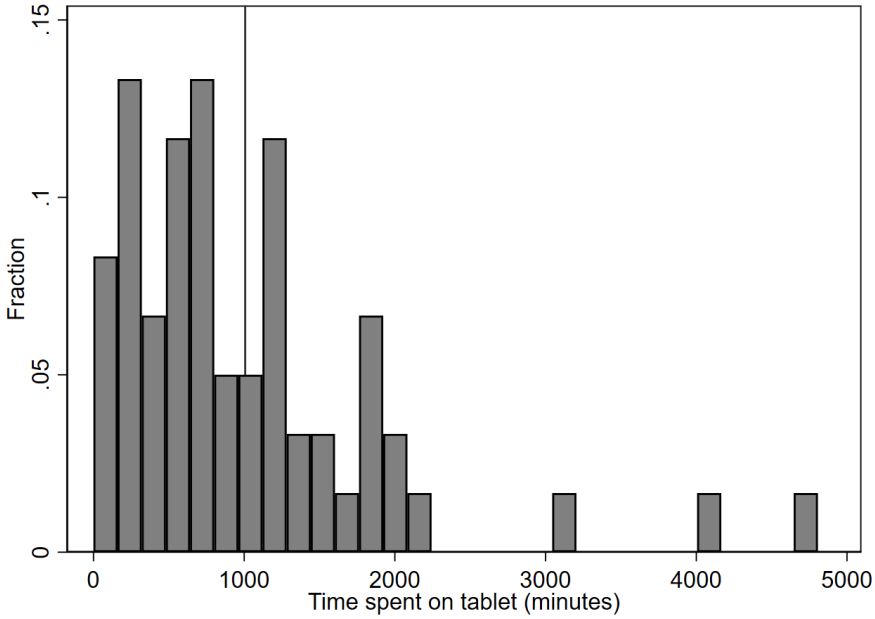


Figure A.2: Histogram of total time spent in term 3 engaging with content on the tablets



Note. Histogram of the total time (in minutes) that teachers in the virtual arm accessed the tablets during the third term of the third year of the program. The line indicates the mean of 1006 minutes (16h45m). The median is 763 minutes (12.7 hours). This excludes time spent on the tablet during training.

Figure A.3: Histogram of number of competitions a teacher entered

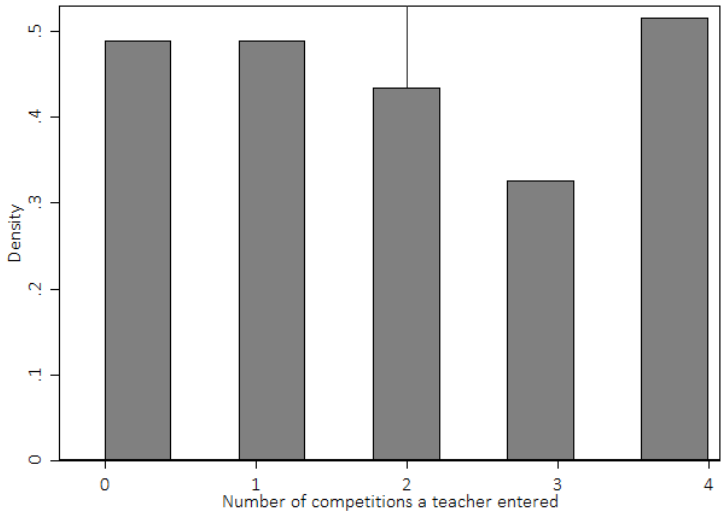
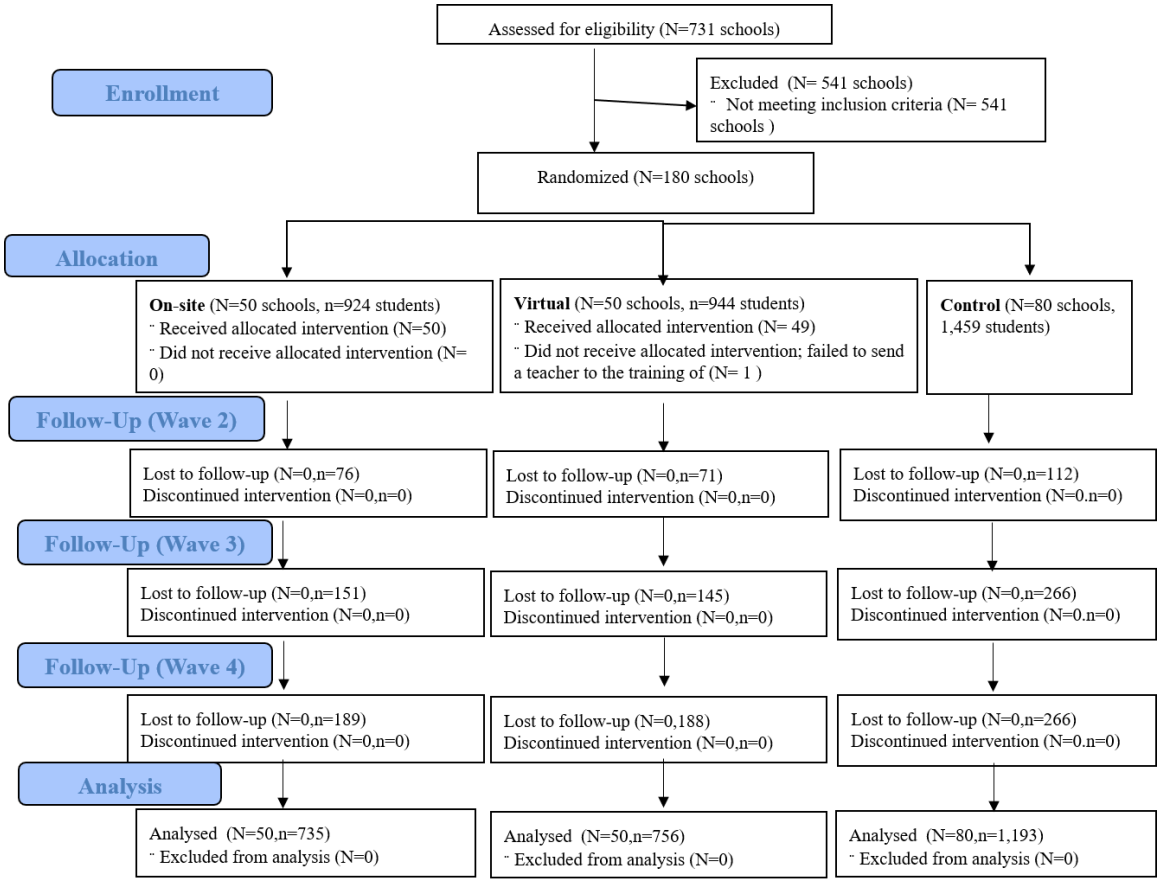


Figure A.4: Summary of attendance at training sessions

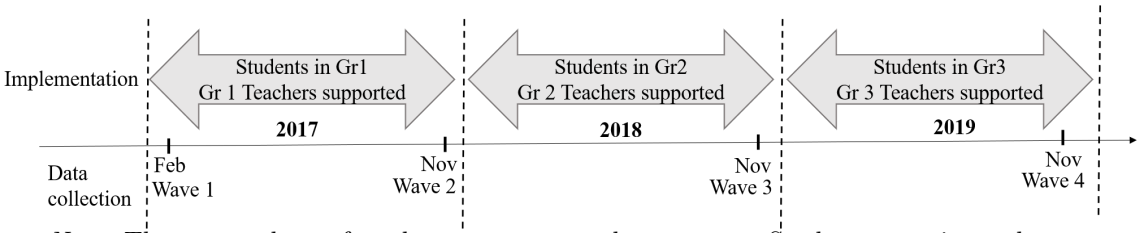
		Total no. teachers	No. teachers trained	No. of SMTs at training
TERM 1	On-site coaching	86	83 (97%)	38 (76%)
	Virtual coaching	85	84 (98%)	31 (63%)
TERM 2	On-site coaching	86	85 (99%)	85 (99%)
	Virtual coaching	83	83 (100%)	25 (51%)
TERM 3	On-site coaching	86	85 (99%)	36 (72%)
	Virtual coaching	82	82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

Figure A.5: Consort diagram



Note. N refers to number of schools; n refers to number of students

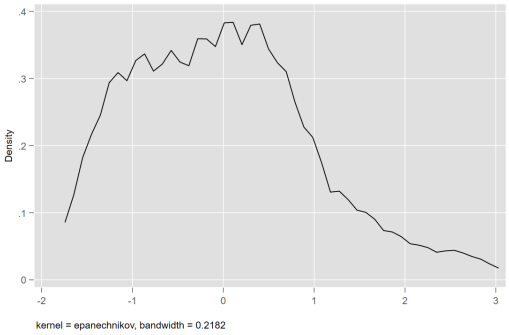
Figure A.6: Timeline



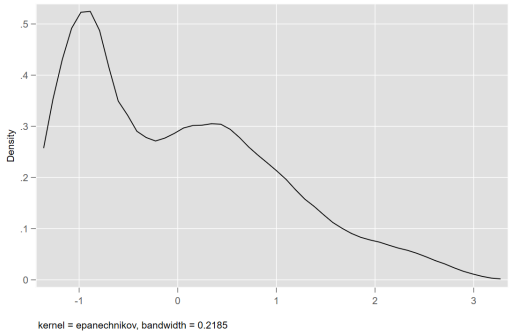
Note. The same cohort of students were assessed every year. Students were in grade one in 2017 (n=3,327). Grade one teachers were surveyed in the first two waves (n=306), grade two teachers in the third wave (n=301), and grade three teachers in the third wave (n=296).

Figure A.7: Kernel Density Plots of English and Home Language Literacy Scores

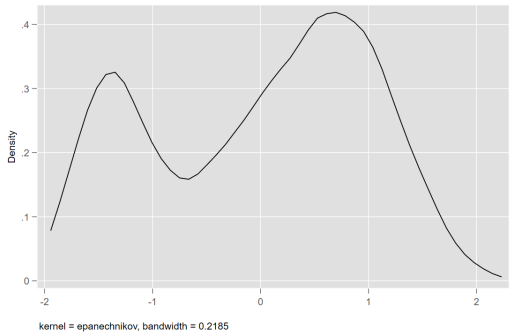
(a) English Oral Proficiency



(b) English Reading Proficiency



(c) Home Language Reading Proficiency



Note. Variables are z-scores of indices constructed using principal components. The English oral language proficiency index is constructed using the English expressive vocabulary task and the English listening comprehension task. The English reading proficiency index is constructed using the English word recognition, English oral reading fluency, English reading comprehension and English written comprehension subtasks. Home language reading fluency is constructed using the letter recognition, oral reading fluency, reading comprehension and written comprehension subtasks. Data is restricted to the control group.

Figure A.8: Percentage of students in the control group who could not read a single word

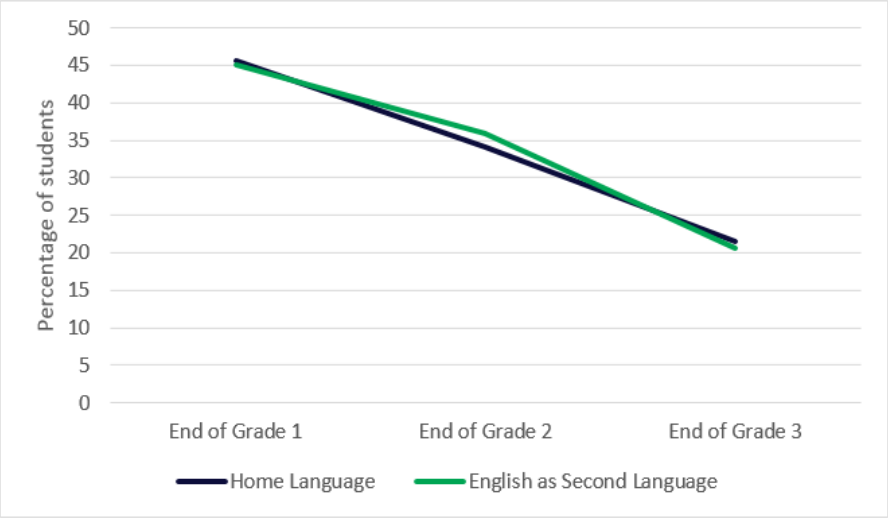


Figure A.9: Cumulative Density Function of Reading Proficiency score, by treatment status

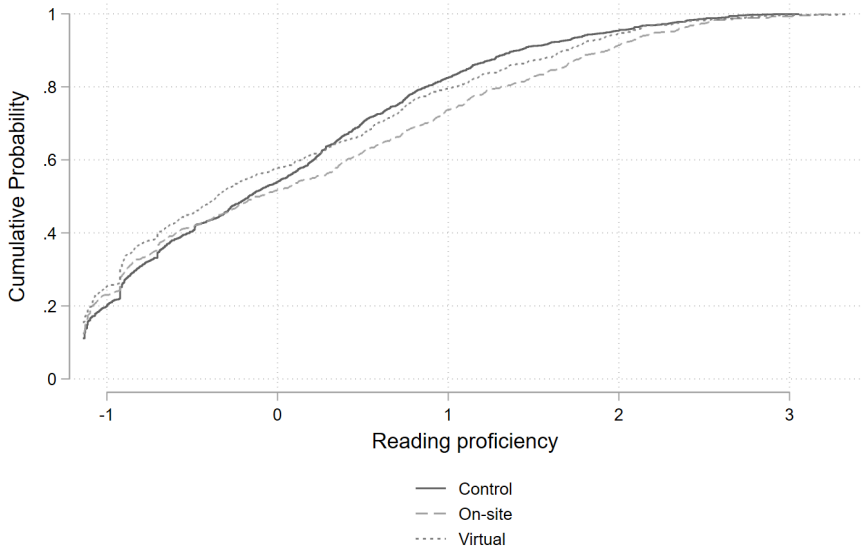
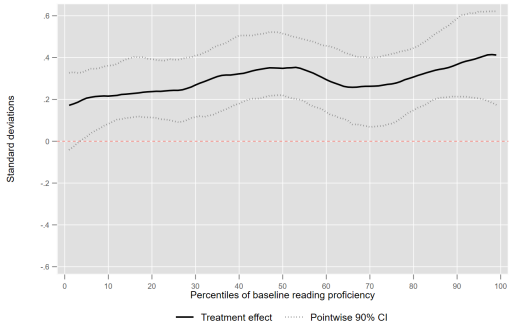
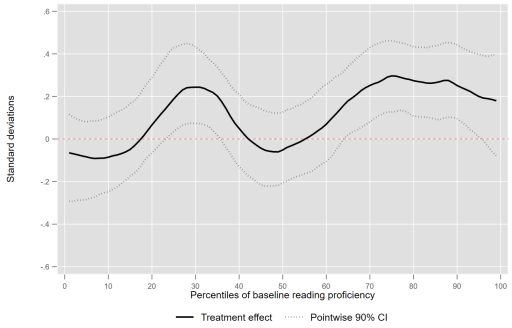


Figure A.10: Interaction with student learning at baseline— English Oral Proficiency

(a) On-site



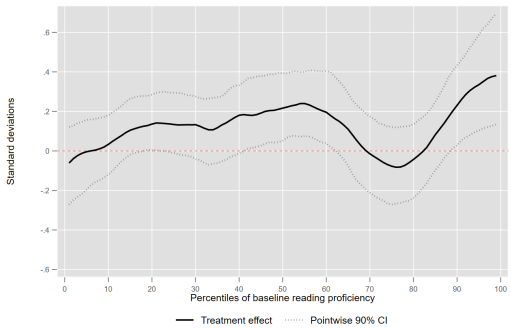
(b) Virtual



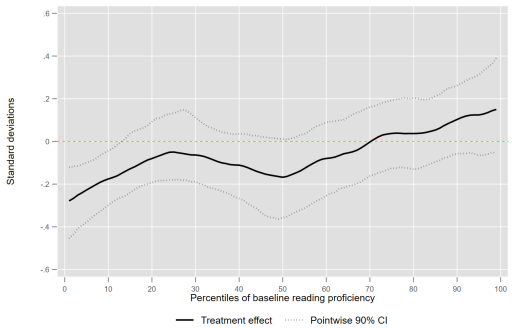
Note. The treatment impacts in Panels (a) and (b) are constructed in four steps. First, we construct a value-added measure of reading proficiency by subtracting the predicted score from the actual score given the set of additional controls in equation 1: $\tilde{y}_{icsb1} = y_{icsb1} - \hat{X_{isb0}}'\Gamma$. Second, we estimate a local polynomial regression of $\tilde{y}_{icsb1}$ on the percentile rank of the aggregate index of baseline learning separately for each treatment arm and the control. Third, we calculate the treatment impact by subtracting the fitted values of each treatment from the fitted values of the control, at each percentile of student baseline performance. Fourth, we construct pointwise 90 percent confidence intervals from a percentile bootstrap with 500 iterations, clustering at the school level and stratifying by randomization strata.

Figure A.11: Interaction with student learning at baseline— English Reading Proficiency

(a) On-site

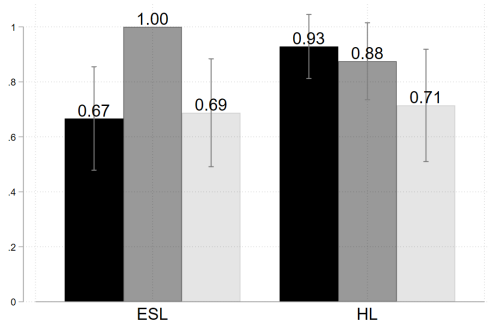


(b) Virtual



Note. See Figure A.10.

Figure A.12: Teacher reported satisfaction with curriculum coverage



Note. Results from teacher questionnaire administered to 51 teachers who participated in the classroom observations. Moving from left to right, the bars indicate the average in the control, on-site, and virtual arms respectively. Lines show 90 percent confidence intervals, with standard errors clustered at the school level.

Table A.1: Required weekly frequency of implementing different learning exercises, by grade

Type of activity	**Grade one**	**Grade two**	**Grade three**
Language use	None	None	Once
Shared reading	Five times	Twice	Twice
Phonemic awareness and phonics	Four times	Three time	Three times
Writing	Once	Twice	Four times
Group-guided reading	None	Five times	Five times

Table A.2: Difference between the on-site and virtual coaching interventions

	On-site	**Virtual**
Lesson plans	Paper-based	Electronic
Media content		Training videos, sound clips, example exercises
Coaching	In person, monthly	Calls every two weeks, weekly text messages, competitions
Training	2-day initial training	3-day initial training

Note: The interventions shared the following features: the service provider, the curriculum, content of the lesson plans, content of the training, 1-day training at the start of each term and additional learning aids such as reading books, posters, flashcards and writing frames.

Table A.3: Subtasks administered to students, by language and wave of data collection

	Start—Year 1		End—Year 1		End—Year 2		End—Year 3	
	HL	ESL	HL	ESL	HL	ESL	HL	ESL
Oral Proficiency								
Receptive Vocabulary		x		x		x		
Expressive Vocabulary	x	x	x	x		x		x
Listening Comprehension	x			x		x		x
Reading Proficiency								
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							
Reading Fluency					x	x	x	x
Reading Comprehension					x	x	x	x
Written Comprehension							x	x

Table A.4: Inter-rater reliability

Variable	(1) Original Mean/SE	(2) Retest Mean/SE	T-test Difference (1)-(2)	Correlation coefficient
HL Oral Reading Fluency	22.727	24.051	-1.324	0.93
	(1.020)	(1.083)		
HL Comprehension	2.295	2.327	-0.032	0.85
	(0.106)	(0.106)		
English Oral Reading Fluency	28.721	28.978	-0.257	0.92
	(1.737)	(1.778)		
English Reading Comprehension	1.083	1.270	-0.187	0.80
	(0.083)	(0.092)		
N	315	315		

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

Table A.5: Balance: School Characteristics

Variable	(1) Control Mean/SE	(2) On-site Mean/SE	(3) Virtual Mean/SE	(4) Total Mean/SE	T-test Difference (1)-(2)	(1)-(3)
Rural	0.738	0.760	0.740	0.744	-0.022	-0.002
	(0.050)	(0.061)	(0.063)	(0.033)		
Bottom quintile	0.537	0.560	0.520	0.539	-0.023	0.018
	(0.056)	(0.071)	(0.071)	(0.037)		
N	80	50	50	180		
F-test of joint significance (p-value)					0.936	0.980
F-test, number of observations					130	130

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.6: Balance: Teacher characteristics

Variable	(1) On-site Mean/SE	(2) Control Mean/SE	(3) Virtual Mean/SE	(4) Total Mean/SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)
At least bachelors	0.695	0.704	0.705	0.702	-0.009	-0.010
	(0.050)	(0.042)	(0.056)	(0.028)		
Class size	44.634	44.244	39.449	43.085	0.390	5.185**
	(1.977)	(1.144)	(1.474)	(0.872)		
Age	46.793	48.785	46.910	47.736	-1.993	-0.118
	(1.141)	(0.804)	(1.294)	(0.599)		
Female	0.976	0.963	0.974	0.969	0.013	0.001
	(0.017)	(0.016)	(0.018)	(0.010)		
Years at school	16.415	18.156	17.471	17.491	-1.742	-1.056
	(1.276)	(0.876)	(1.335)	(0.639)		
N	82	135	78	295		
Clusters	50	80	50	180		
F-test of joint significance (p-value)					0.645	0.341
F-test, number of observations					217	160

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable NatEmis. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.7: Balance: Student characteristics

Variable	(1) Control Mean/SE	(2) On-site Mean/SE	(3) Virtual Mean/SE	(4) Total Mean/SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)
Age	6.087	6.113	6.140	6.109	-0.026	-0.053
	(0.035)	(0.040)	(0.050)	(0.024)		
Male	0.534	0.544	0.550	0.541	-0.010	-0.016
	(0.013)	(0.015)	(0.017)	(0.009)		
Zulu	0.307	0.291	0.267	0.291	0.016	0.040
	(0.053)	(0.066)	(0.063)	(0.034)		
Naming Animals in HL	7.155	7.310	7.501	7.296	-0.155	-0.346*
	(0.127)	(0.155)	(0.154)	(0.083)		
Word recall	9.981	9.953	10.081	10.002	0.028	-0.099
	(0.084)	(0.093)	(0.092)	(0.052)		
Nonword recall	4.208	4.179	4.237	4.208	0.029	-0.030
	(0.049)	(0.052)	(0.082)	(0.035)		
Phoneme isolation	1.129	1.037	1.161	1.112	0.092	-0.032
	(0.087)	(0.092)	(0.107)	(0.055)		
Story comprehension	2.179	2.154	2.263	2.196	0.025	-0.084
	(0.045)	(0.050)	(0.047)	(0.028)		
Letters sound correct	6.978	6.784	7.019	6.936	0.194	-0.041
	(0.447)	(0.590)	(0.610)	(0.307)		
Words Correct	0.387	0.347	0.510	0.411	0.039	-0.123
	(0.096)	(0.103)	(0.148)	(0.066)		
Sentence Correct	0.051	0.027	0.034	0.040	0.024	0.018
	(0.012)	(0.011)	(0.012)	(0.007)		
Visual Perception	1.460	1.597	1.651	1.552	-0.137	-0.192
	(0.082)	(0.111)	(0.109)	(0.057)		
English Items	0.836	0.789	0.839	0.824	0.047	-0.003
	(0.044)	(0.063)	(0.045)	(0.029)		
N	1459	924	944	3327		
Clusters	80	50	50	180		
F-test of joint significance (p-value)					0.884	0.230
F-test, number of observations					2383	2403

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable NatEmis. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.8: Balance: Classroom Observation Sample

Variable	(1) Control Mean/SE	(2) On-site Mean/SE	(3) Virtual Mean/SE	(4) Total Mean/SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)
School level						
Rural	0.778	0.765	0.722	0.755	0.013	0.056
	(0.101)	(0.106)	(0.109)	(0.060)		
Bottom quintile	0.556	0.647	0.556	0.585	-0.092	0.000
	(0.121)	(0.119)	(0.121)	(0.068)		
N	18	17	18	53		
Teacher level						
Class size	42.794	47.192	36.600	42.000	-4.398	6.194**
	(2.265)	(4.824)	(1.388)	(1.785)		
Age	48.059	46.885	46.933	47.344	1.174	1.125
	(1.889)	(2.211)	(2.480)	(1.244)		
Female	0.941	1.000	0.967	0.967	-0.059	-0.025
	(0.036)	(0.000)	(0.034)	(0.018)		
Years at school	19.588	18.154	17.590	18.508	1.434	1.998
	(1.693)	(2.528)	(2.507)	(1.260)		
N	34	26	30	90		
Student level						
Age	6.016	6.142	6.148	6.103	-0.126	-0.132
	(0.071)	(0.047)	(0.084)	(0.040)		
Male	0.521	0.558	0.537	0.539	-0.038	-0.016
	(0.030)	(0.031)	(0.030)	(0.017)		
Baseline Reading	-0.005	-0.124	0.133	0.004	0.119	-0.138
	(0.117)	(0.091)	(0.064)	(0.054)		
N	315	317	337	969		

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.9: Descriptive statistics of each assessment subtask administered to students at baseline

	Mean	SD	Min	Max	p10	p25	p50	p75	p90
Home Language									
Letter Recog.	44.00	22.44	0.00	110.00	13.00	28.00	44.50	60.00	72.00
Read. Fluency	21.99	17.76	0.00	58.00	0.00	1.00	23.00	36.00	47.00
Read. Compr.	2.30	1.91	0.00	5.00	0.00	0.00	3.00	4.00	5.00
W. Compr.	2.30	1.91	0.00	6.00	0.00	0.00	2.00	4.00	5.00
English									
Word Recog.	23.78	21.86	0.00	99.00	0.00	1.00	22.00	40.00	55.00
Read. Fluency	28.29	30.44	0.00	126.00	0.00	0.00	19.00	49.00	72.00
Read. Compr.	1.07	1.46	0.00	5.00	0.00	0.00	0.00	2.00	4.00
Vocab.	3.29	1.78	0.00	6.00	1.00	2.00	3.00	5.00	6.00
L. Compr.	0.99	1.07	0.00	4.00	0.00	0.00	1.00	1.00	3.00
W. Compr.	1.43	1.25	0.00	4.00	0.00	0.00	1.00	2.00	3.00

Table A.10: Attrition

	(1) Attrite	(2) Age	(3) Male	(4) isiZulu	(5) Learning
On-site coach	0.025	0.032	0.014	-0.020	-0.003
	(0.021)	(0.054)	(0.021)	(0.078)	(0.071)
Virtual coach	0.016	0.023	0.013	-0.049	0.110
	(0.023)	(0.059)	(0.025)	(0.077)	(0.071)
Attrite		-0.031	-0.004	0.021	-0.021
		(0.052)	(0.035)	(0.033)	(0.067)
Attrite x On-site		0.022	-0.016	0.004	-0.064
		(0.078)	(0.053)	(0.061)	(0.099)
Attrite x Virtual		0.139	0.010	0.061	0.009
		(0.096)	(0.062)	(0.067)	(0.112)
Mean attrition in control	0.18				
Observations	3327	3327	3327	3327	3327
R-squared	0.004	0.016	0.002	0.145	0.023

Notes: Standard errors are clustered at the school level. Estimates include strata fixed effects.

Table A.11: Balance after attrition

Variable	(1) Control Mean/SE	(2) On-site Mean/SE	(3) Virtual Mean/SE	(4) Total Mean/SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)
Age	6.093	6.112	6.116	6.105	-0.019	-0.023
	(0.036)	(0.042)	(0.052)	(0.024)		
Male	0.535	0.548	0.549	0.542	-0.014	-0.014
	(0.014)	(0.016)	(0.020)	(0.009)		
Zulu	0.303	0.288	0.247	0.284	0.015	0.056
	(0.053)	(0.067)	(0.062)	(0.035)		
Naming Animals in HL	7.231	7.329	7.508	7.336	-0.099	-0.277
	(0.135)	(0.164)	(0.161)	(0.087)		
Word recall	9.999	9.948	10.053	10.000	0.051	-0.054
	(0.089)	(0.107)	(0.093)	(0.055)		
Nonword recall	4.206	4.188	4.280	4.222	0.018	-0.074
	(0.051)	(0.059)	(0.075)	(0.035)		
Phoneme isolation	1.110	1.097	1.180	1.126	0.013	-0.070
	(0.084)	(0.099)	(0.114)	(0.056)		
Story comprehension	2.191	2.161	2.228	2.193	0.031	-0.036
	(0.048)	(0.059)	(0.048)	(0.030)		
Letters sounds correct	6.983	7.006	7.101	7.023	-0.023	-0.118
	(0.442)	(0.633)	(0.632)	(0.315)		
Words Correct	0.362	0.362	0.496	0.400	0.000	-0.134
	(0.093)	(0.116)	(0.150)	(0.067)		
Sentences Correct	0.042	0.030	0.038	0.038	0.012	0.004
	(0.012)	(0.014)	(0.014)	(0.008)		
Visual Perception	1.495	1.537	1.648	1.550	-0.042	-0.153
	(0.091)	(0.106)	(0.115)	(0.059)		
English Items	0.828	0.777	0.819	0.811	0.051	0.009
	(0.047)	(0.055)	(0.051)	(0.029)		
N	1193	735	756	2684		
Clusters	80	50	50	180		
F-test of joint significance (p-value)					0.998	0.707
F-test, number of observations					1928	1949

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable NatEmis. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.12: Probability of reaching grade three

	(1) Attrite	(2) Age	(3) Male	(4) isiZulu	(5) Learning
On-site coach	-0.053	0.082	-0.034	-0.011	-0.079
	(0.026)	(0.073)	(0.036)	(0.085)	(0.080)
Virtual coach	-0.054	0.055	0.016	-0.043	0.091
	(0.028)	(0.086)	(0.034)	(0.086)	(0.091)
Grade 3		0.094	-0.118	-0.034	0.313
		(0.054)	(0.028)	(0.031)	(0.054)
Grade 3 x On-site		-0.065	0.062	-0.015	0.127
		(0.075)	(0.044)	(0.048)	(0.100)
Grade 3 x Virtual		0.001	-0.011	0.008	0.061
		(0.084)	(0.045)	(0.053)	(0.091)
Prop grade 3	0.68				
Observations	3327	3327	3327	3327	3327
R-squared	0.005	0.018	0.013	0.145	0.054

Notes: Standard errors are clustered at the school level. Estimates include strata fixed effects.

Table A.13: Quality of implementation

	Received training			Graded readers		Use graded readers		Use lesson plans		Classroom quality		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Grade 1	Grade 2	Grade 3	Grade 2	Grade 3	Grade 2	Grade 3	Grade 2	Grade 3	Grade 1	Grade 2	Grade
On-site	0.231	0.258	0.262	0.366	0.297	0.365	0.276	0.382	0.113	0.503	0.868	0.894
	(0.055)	(0.046)	(0.058)	(0.054)	(0.053)	(0.055)	(0.055)	(0.066)	(0.022)	(0.172)	(0.148)	(0.150
Virtual	0.256	0.208	0.234	0.344	0.314	0.342	0.338	0.255	0.140	0.368	0.790	0.755
	(0.053)	(0.051)	(0.064)	(0.058)	(0.054)	(0.059)	(0.053)	(0.073)	(0.021)	(0.175)	(0.165)	(0.142
Observations	306	301	292	301	279	301	296	301	904	306	301	292
R-squared	0.144	0.137	0.106	0.231	0.207	0.219	0.186	0.153	0.025	0.104	0.200	0.182
Control	0.745	0.733	0.639	0.641	0.683	0.634	0.637	0.542	0.174	-0.000	-0.000	0.000
On-site	0.941	0.977	0.914	0.943	0.961	0.931	0.892	0.862	0.263	0.642	0.942	0.960
Virtual	0.964	0.928	0.923	0.928	0.961	0.916	0.949	0.735	0.276	0.547	0.841	0.736

Notes: Each column represents a separate regression, based on the teacher surveys conducted on grade 1, 2, and 3 teachers, in 2017, 20 and 2019 respectively. Standard errors are clustered at the school level. Estimates include strata fixed effects. "Received training" i binary variable equal to one if the teachers indicated that they received professional in-service teacher training for ESL in the year of da collection; "Graded readers" is a binary variable equal to one if the teacher has graded readers for ESL in her classroom; "Use grad readers" indicates if the teacher reports to use the graded reader; "Use lesson plans" is a binary variable equal to one if the teacher u lesson plans provided by an NGO. "Classroom quality" is a Kling index, standardized to a control mean of zero and standard deviat of one, consisting of the following indicators collected during classroom observations: the availability of a reading corner, the number storybooks, the quality of posters, and the quality of flashcards.

Table A.14: Robustness Check—Inverse Probability Weights

	(1) Oral proficiency	(2) Reading proficiency
On-site coach	0.288	0.121
	(0.065)	(0.069)
Virtual coach	0.118	-0.054
	(0.071)	(0.068)
Observations	2684	2632
R-squared	0.292	0.299
F-test	0.028	0.020

Notes. See table 1. Each regression is weighted by the inverse of the predicted probability of a student attriting, based on observed characteristics. The probability of attriting is estimated using a probit model, with the following predictors: students' scores on the baseline sub-tasks, gender, age, and district.

Table A.15: Robustness check—Lee bounds

	Oral proficiency		Reading proficiency	
	(1) Lower	(2) Upper	(3) Lower	(4) Upper
On-site coach	0.289	0.364	0.114	0.180
	(0.068)	(0.066)	(0.068)	(0.066)
Virtual coach	0.099	0.180	-0.062	0.010
	(0.072)	(0.071)	(0.069)	(0.066)
Control mean	0.037	-0.071	0.028	-0.077
Observations	2656	2653	2604	2597
R-squared	0.289	0.292	0.295	0.294
F-Test	0.021	0.025	0.018	0.022

Notes. Each column represents a separate regression, estimated using equation 1. Columns (2) and (4) show the upper Lee (2009) bounds of the estimated treatment effect, trimming the top 2.39 percent of students in the control. Columns (1) and (3) show the lower Lee bound of the estimated treatment effect, trimming the bottom 2.39 percent of students in the control.

Table A.16: Robustness Check: No student-level weights

	(1) Oral proficiency	(2) Reading proficiency
On-site coach	0.288	0.121
	(0.065)	(0.069)
Virtual coach	0.118	-0.054
	(0.071)	(0.068)
Observations	2684	2632
R-squared	0.292	0.299
F-test	0.028	0.020

Notes. See table 1. Regressions do not include any weights.

Table A.17: Extended English Vocabulary Assessment vs Original Instrument

	(1) Original	(2) Extended
On-site	3.945	5.032
	(2.698)	(2.069)
Virtual	-1.329	2.858
	(2.682)	(2.654)
Control mean	21.139	20.685
Observations	315	315
R-squared	0.398	0.519
F-test	0.085	0.329

Notes. Standard errors are clustered at the school level. Estimates include strata fixed effects.Column (1) shows the treatment effects using the original instrument, but restricted to the sample of students who also participated in the expanded vocabulary test. The second column shows the treatment effects using the expanded English vocabulary test.

Table A.18: Impact of on-site and virtual coaching on English oral proficiency, by year and subtask

	Vocabulary			Comprehension		
	(1)	(2)	(3)	(4)	(5)	(6)
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
On-site coach	0.442	0.151	0.420	0.030	0.041	0.074
	(0.056)	(0.082)	(0.105)	(0.011)	(0.016)	(0.017)
Virtual coach	0.468	0.159	0.147	0.031	0.006	0.032
	(0.053)	(0.074)	(0.118)	(0.010)	(0.015)	(0.018)
Control mean	0.484	4.175	3.120	0.058	0.205	0.216
F-Test	0.662	0.928	0.032	0.933	0.034	0.036
Observations	3061	2763	2684	3063	2762	2684
R-squared	0.283	0.249	0.266	0.138	0.268	0.231

Notes:. See table 1. Data comes from assessments conducted to the same students at the end of the 1^{st}, 2^{nd} and 3^{rd} years of the evaluation, respectively. "Vocabulary" is the number of English words understood by the student; "Comprehension" is the proportion of questions that the student answered correctly in an listening comprehension test. The assessments are adapted every year, so are not directly comparable across time.

Table A.19: Impact of on-site and virtual coaching on English reading proficiency, by year and subtask

	Year 2			Year 3			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Word recog.	Read. fluency	Read. compr.	Word recog.	Read. fluency	Read. compr.	Written compr.
On-site coach	1.103	1.174	0.002	2.660	2.458	0.060	0.016
	(1.324)	(1.356)	(0.010)	(1.360)	(1.909)	(0.019)	(0.020)
Virtual coach	-1.576	-0.878	-0.012	-1.199	-0.818	0.016	-0.019
	(1.251)	(1.438)	(0.009)	(1.357)	(1.902)	(0.018)	(0.019)
Control mean	17.016	18.855	0.145	23.121	27.255	0.191	0.355
F-Test	0.071	0.198	0.176	0.009	0.108	0.037	0.112
Observations	2764	2764	2765	2684	2684	2684	2632
R-squared	0.270	0.294	0.253	0.254	0.265	0.264	0.218

Notes:. See tables 1. Data comes from students assessments conducted at the end of the 2^{nd} and 3^{rd} years of the evaluation, respectively. The assessments are adapted every year, so are not directly comparable across time.

Table A.20: Interaction with baseline learning

	(1) Oral Proficiency	(2) Reading Proficiency
On-site coach	0.335	0.141
	(0.072)	(0.068)
Virtual coach	0.112	-0.063
	(0.072)	(0.067)
On-site x Baseline learning	0.067	0.086
	(0.051)	(0.047)
Virtual x Baseline learning	0.048	0.094
	(0.061)	(0.052)
Control mean	0.000	0.000
Observations	2684	2632
R-squared	0.264	0.267
F-test	0.009	0.009

Notes. Standard errors are clustered at the school level. Estimates include strata and enumerator fixed effects, and controls for student gender, age, and district. "Baseline learning" is a index of baseline learning proficiency, constructed using principal component analysis, and standardized to have a control mean of zero and standard deviation of one.

Table A.21: Correct frequency of learning activities (self-reported)

	(1) Phonics sound	(2) Phonics lesson	(3) Group-guided reading	(4) Shared Reading
On-site	0.209	0.469	0.110	0.288
	(0.072)	(0.067)	(0.067)	(0.071)
Virtual	0.194	0.240	0.156	0.280
	(0.077)	(0.079)	(0.067)	(0.076)
Control mean	0.149	0.074	0.213	0.136
Observations	296	296	296	296
R-squared	0.084	0.172	0.066	0.092
F-test	0.861	0.007	0.516	0.913

Notes. Each column is a separate regression, estimated using equation 1. Data comes from teacher surveys administered to all grade three teachers. Estimates include strata fixed effects, and standard errors are clustered at the school level. The final row reports the p-value of the F-test of equality of coefficients. Each outcome is a binary variable indicating whether the teacher's self-reported frequency of implementing a teaching technique is the same as the required weekly frequency. See table A.1 for the required weekly frequencies.

Table A.22: Learning activities observed during the home language lesson

	(1) Language Phonics	(2) Shared reading	(3) Group-guided reading	(4) Pupil reads individually	(5) Writing
On-site	-0.294	0.256	0.169	0.269	0.030
	(0.127)	(0.188)	(0.185)	(0.120)	(0.178)
Virtual	-0.056	0.005	0.072	0.249	0.290
	(0.113)	(0.180)	(0.187)	(0.133)	(0.149)
Observations	44	44	44	44	44
R-squared	0.395	0.252	0.187	0.346	0.283
Control mean	0.882	0.529	0.235	0.059	0.647
F-test	0.116	0.202	0.592	0.892	0.118

Notes. Data comes from classroom observations conducted when the teacher was teaching the home language lesson, restricted to teachers who were teaching both English and Home Language on the day of classroom observations. Outcomes are dummy variables equal to one if the respective teaching activities took place at least once during the full duration of the lesson.

Table A.23: Mediation Analysis: Group-guided reading

	Oral Proficiency		Reading Proficiency	
	(1) Outcome	(2) Demediated	(3) Outcome	(4) Demediated
On-site coach	0.340	0.263	0.175	0.072
	(0.079)	(0.079)	(0.076)	(0.074)
Virtual coach	0.089	0.057	-0.038	-0.079
	(0.077)	(0.075)	(0.075)	(0.072)
Observations	2027	2027	2000	2000
R-squared	0.268	0.262	0.271	0.269

Notes. Each column represents a separate regression, using same set of controls as in table 1. In columns (2) and (4) the outcome variables are demediated using the sequential g-estimation (Acharya et al., 2016). See section A.4 for an explanation of methods, and choice of additional controls of potential confounders in the first stage of demediating the outcome.

Table A.24: Interaction between teacher age and use of lesson plan

	(1)	(2)
Virtual	0.262	0.096
	(0.197)	(0.068)
Age	-0.002	
	(0.003)	
Age x Virtual	-0.004	
	(0.004)	
Old (> 55)		0.098
		(0.086)
Old (> 55) x Virtual		-0.192
		(0.133)
On-site mean	0.807	0.807
Observations	161	161
R-squared	0.062	0.057

Notes. Each column represents a separate regression. Regression estimates also include strata fixed effects. The outcome is a binary variable equal to one if a teacher reports to use a lesson plans provided by an NGO. The variable "Old" is equal to one if a teacher is older than 55. Data is at a teacher level and restricted to the two treatment arms. Standard errors are clustered at the school level. Estimates include strata fixed effects.

Table A.25: Breakdown of costs (per student per year, USD)

	USD		**% On-site total**	
	On-site	**Virtual**	**On-site**	**Virtual**
Support personnel	9.8	11.7	15%	18%
Learning aids	13.3	13.6	20%	21%
Training	8.5	10.8	13%	16%
Reading coaches	24.9	7.0	38%	11%
Travel	8.5	0.9	13%	1%
Lesson plans	0.9	0.0	1%	
Tablets		2.6		4%
Data and communication		2.8		4%
App maintenance		2.5		4%
Total	65.8	51.9	100%	81%

Notes. Costs are per student per year in USD, taking a ZAR:USD exchange rate of 14:1. The right-hand columns report costs as a proportion of the total costs for the on-site coaching program. Costs of developing and piloting the program, and purchasing the tablets are not included. Tablet depreciation costs are included and calculated by the sum of year's digits method, assuming a lifespan of seven years. $3,550$ children were supported by each program.