RISE Working Paper 20/050
September 2020

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

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

Abstract

We experimentally compare on-site with virtual coaching of South African teachers. After three years, on-site coaching improved students' English oral language and reading proficiency by 0.31 and 0.13 SD, respectively. Virtual coaching improved English oral language proficiency (0.12 SD), had no impact on English reading proficiency, and an unintended negative effect on home language literacy. Classroom observations show that on-site coaching improved teaching practice and that virtual coaching led to larger crowding-out of home language teaching time. Implementation and survey data suggest that the use of technology did not preclude effectiveness, but rather that in-person contact enabled more accountability and support.

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

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

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

Janeli Kotzé
Department of Basic Education, Government of
South Africa

Nompumelelo Mohohlwane
Department of Basic Education, Government of
South Africa

Stephen Taylor
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. 2020. Can Virtual Replace In-person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning in South Africa. RISE Working Paper Series. 20/050. https://doi.org/10.35489/BSGRISEWP_ 2020/050.

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

What is the most cost effective way to improve teaching quality at scale? This question is perhaps one of the most important challenges facing education systems in developing countries today, where the levels of learning in schools are low, teachers work in highly challenging environments, and teacher mastery of content and pedagogical skills is weak (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. However, governments have limited discretionary budgets, with the majority of funding allocated to salaries and infrastructure, and so typically implement low-cost teacher training models, often conducted outside the classroom, with a cascading design (”train the trainers”), and with limited ongoing support. Many have questioned the efficacy of such programs (Popova et al., 2016). This challenge will only increase as schools start to re-open after closures due to the COVID-19 pandemic, since teachers will face even more academically diverse classrooms and more emotionally strenuous environments. Yet government education budgets will be even more constrained than before.

Information communications technology (ICT) presents the possibility of lowering the cost of providing support to teachers at scale. An expert trainer or coach could reach more teachers virtually, compared to in-person visits or training. This would not only reduce transport and salary costs, but could address potential binding human capital constraints in finding enough high-quality trainers. New technology might also improve the nature of the support materials provided to teachers, such as additional instructional videos or lesson plans, which are available to teachers at any time not only at training or when a coach is present to help.

However, there are also reasons why virtual teacher professional development may be less effective. Teachers might struggle to adapt to using new technology, and require substantial training up-front to use the technology. Moreover, face-to-face engagement might be necessary to build a relationship of trust between the teacher and the trainer or coach, which allows the teacher to be vulnerable and discuss ways to improve her teaching. A lack of face-to-face engagement could also lead to less accountability, all of which means that such an intervention relies on a greater degree of self-motivation from teachers.

Can virtual instruction replace in-person instruction? 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. This evaluation is the second in a set of randomized evaluations run by South Africa’s Department of Basic Education, called the Early Grade Reading Studies (EGRS).[1] The first EGRS compared the impact of monthly on-site coaching to that of a traditional teacher training program (two days residential training

[1] See Fleisch (2018) for an overview, as well as background and rationale for the development of the programs.

twice a year). It was found that on-site coaching improved home language reading outcomes by 0.24 standard deviations, whereas the training intervention did so by only 0.12 standard deviations (Cilliers et al., 2019). Despite this promising evidence in support of coaching, the aforementioned questions about cost and scale are being raised at a time of great fiscal constraints in the country.[2] The second EGRS, the focus of this paper, was designed with this question of cost and scalability in mind.

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.[3] In both programs teachers received the same learning materials and training at the start of the program, and the curriculum and content of lesson plans were the same. 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, 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 sub-set of 53 schools at the end of the third year.

We highlight four 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 had statistically significant positive impacts on both English oral language proficiency (0.31 standard deviations) and English reading proficiency (0.13 standard deviations). In contrast, the virtual coaching program was far less effective at improving English oral language proficiency (0.12 standard deviations), and had no statistically detectable impact on reading proficiency skills. The difference in effect sizes between on-site and virtual coaching is statistically significant at a 5 percent level, for both outcomes. The on-site coaching program is about 23 percent more expensive than the virtual coaching program, but the cost-effectiveness analysis shows that it

[2] At the time of writing, the Department of Basic Education had commissioned a costing analysis to make recommendations about how to implement aspects of the successful EGRS program on a larger scale.

[3] 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 teacher training that is organised and presented by the district or the province.

is still more cost-effective. This finding is in contrast with the initial evaluation results after just one year of implementation, which found comparable impacts between on-site and virtual coaching on English oral language proficiency (Kotze et al., 2019).[4]

Second, virtual coaching *reduced* home language reading proficiency by 0.19 standard deviations and caused a reallocation of time inputs from home language (HL) to ESL instruction. The difference between the on-site coaching and virtual coaching treatment groups in HL language reading proficiency is statistically significant at the 10 percent level. 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: the proportion of teachers who dedicate less than the minimum requirement set by the curriculum almost doubled.

Third, classroom observations reveal that the on-site coaching induced larger gains in teacher productivity, relative to virtual coaching. 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 and higher order pedagogical skills, such as group-guided reading and independent reading, were better and more frequently implemented by teachers who had received on-site coaching.

Fourth, we are able to rule out differences in fidelity of program implementation or the format of the lesson plans (electronic versus paper-based) as the reason for the virtual coaching program being less effective than on-site coaching program, and therefore we 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 high. Moreover, tablet usage data suggests that technology itself was not the main barrier to program implementation, since almost all the teachers in the virtual arm used tablets and accessed the lesson plans on more than a handful of occasions. Interestingly, tablet usage was better earlier in the term than towards the end of the term, and was highest in the week in which teachers were expected to submit assessment results. This pattern would suggest that the technology itself was not the main barrier to program implementation, but rather the motivation of teachers or their ability to keep pace with the curriculum. Consistent with this interpretation, teachers in the on-site coaching intervention were more likely to mention the coach as someone who holds them accountable and someone who provides pedagogical support. Teachers in the on-site coaching arm were also more likely to report ever being observed teaching or to have seen somebody modeling teaching practice. This suggests that the direct observation and opportunities for feedback available to an on-site coach were ultimately critical to program success.

This study contributes to two strands of literature. First, in terms of teacher professional development, a

[4] We did not measure English reading proficiency at the end of grade one, since the grade one ESL curriculum focuses exclusively on developing language proficiency. Decoding skills are only introduced in grades two and three.

growing body of research has demonstrated the important role that pedagogical coaches can play in improving student learning (Kraft et al., 2018). One of the most effective classes of interventions ever evaluated in developing countries is structured pedagogy programs that combine carefully planned curriculum (often with daily lesson plans), additional learning aids such as reading booklets, and pedagogical coaches (Cilliers et al., 2019; Evans and Popova, 2015; 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, 2018).

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.[5]

Our findings contrast evidence from the United States on the relative effectiveness of on-site vs 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. Similarly, 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 follow-ups. As McEwan (2015) notes, most studies on education interventions show impacts after just one year (or less). Initial gains in language proficiency may not be sustained over time or as students transition to more advanced skills.

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.

[5] 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.

In the 2019 matriculation examinations, for example, 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 (Filmer et al., 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 English is 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. It is also increasingly recognised in a context of decolonisation that there may be intrinsic reasons for more extensive use of home language in education. In practice, most children in South Africa learn in their home language as the main language of instruction during grades one to three and then experience a transition to English as the language of instruction from grade four onward.[6] To ameliorate the language transition learners face in Grade 4, English is introduced as an additional language from Grade 1.

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.[7] Both interventions consist of three interrelated 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. The

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

[7] 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.

content and support materials provided were the same in both interventions and were fully aligned to the official ESL curriculum. This means that students were taught oral language proficiency skills during the English lessons in the first grade, and decoding (i.e. reading skills) was only introduced in the English lesson from the second half of the second grade.[8] 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.

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. GGR gives the teacher an opportunity to provide more individualized feedback to each student, but is a difficult technique to implement, since 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. 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 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 B.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 B.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

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

answered questions over WhatsApp on an ongoing basis.[9] 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.[10] The content was updated quarterly and designed to work offline; connectivity was therefore not a barrier for daily usage. Figure B.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 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 B.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.[11] 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 B.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

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

[10] 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.

[11] 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

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 of students. About 7,600 students benefited from the interventions for the three year period.

We randomly selected 180 public primary schools out of a population of schools that meet the following criteria: (i) only non-fee paying public schools[12], (ii) primary language of instruction is Siswati or isiZulu, and (iii) grade one enrollment is between 30 and 160.[13] We then created 10 strata of similar schools, based on school size, socio-economic status and previous performance in the Annual National Assessments, 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.

3 Mechanisms

Broadly speaking, coaches can play three roles. The first is providing technical 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, and arguably the most important role, is one of a confidante, where the coach builds a trust relationship with teachers that would emotionally prepare teachers for changing their instructional practices. Qualitative work conducted by Alsofrom (2019) found that teachers in this study come into the program with expectations formed through previous experiences. Often these experiences have conditioned teachers to expect negative feedback from observers but without the resources 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. This emotional shift is a critical pre-requisite to teachers accepting the targeted feedback on their instructional practices, and subsequently for their practices to change in a deep and meaningful way. The enactment of these roles looks very different between the on-site coaches and the

[12] 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.

[13] 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

virtual coach and the study aims to evaluate whether both methods can be equally effective.

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

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 a one-on-one setting with the sampled students, whereas the written assessments were administered in a group setting. 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. Figure B.5 provides a summary of the different components of student assessment administered in 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 in the Foundation Phase.[14] The first construct is oral language proficiency as it relates to English vocabulary

[14] The study was registered with the AEA Trial Registry: https://doi.org/10.1257/rct.5148-1.0.

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 questionnaire 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. 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. The document review and classroom observation schedules captured information on the number of written activities that students engaged in, as well as the print richness of the classroom environment. Fieldworkers were asked to look at the classroom and rate the print richness of the environment on a scale from one to four, with four being a very print-rich classroom with high-quality materials.

Three additional evaluation activities were conducted at the end of the third year of implementation, each aimed at providing a different perspective on whether the treatments were successful and the mechanisms which contributed to the success. The first activity entailed re-testing 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 extended 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.3 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, conducted during the third term of the third year of the study. We randomly sampled twenty teachers in each treatment arm —stratifying by the language of instruction in the school (isiZulu or Siswati) and baseline learning outcomes— but 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, 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 (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.

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 B.6 shows kernel density plots of ESL Oral and Reading Proficiency, as well as HL reading proficiency.[15] 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. Note also the bi-modal distribution in the ESL and HL Reading Proficiency. This suggests that there is a large proportion of students who had not yet out of the starting blocks on a path towards reading fluency and comprehension, even after three years of schooling. Indeed, Figure B.7 shows that at the end of grade one, about 46 percent of students in the control could not read a single word correctly in their HL. It is disconcerting that two years later, at the end of Grade three, 22 percent of the grade three students still did not read a single word correctly in HL. The percentage of non-readers in English was about the same (21 percent by the end of grade three), even though the word length in the English language is shorter than in isiZulu and Siswati.

Table A.9, 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 observed characteristics do not predict attrition status, and the treatments do not change the composition of attriters. Table A.10 shows that the sample remains balanced if we exclude the attriters. It is therefore unlikely that attrition would bias the results. Table A.11, 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. Unexpectedly, students in both treatment arms were 5 percentage points *less* likely to reach Grade three. 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:

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

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.[16] Moreover, since attrition was not uniform across schools, we also re-weight each observation based on number of students so 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

[15] See Table A.4 for the descriptive statistics of each assessment instrument administered to students at baseline.

[16] 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.

only include the strata as controls.

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. Figure 1 shows teachers' exposure to key components that were shared across the programs: attending training, receipt of lesson plans (either tablet or paper-based), and receipt of graded reading booklets. Around 90 percent of teachers in the on-site coaching and virtual coaching treatment arms reported that they attended training for ESL in 2019. Moreover, nearly all teachers indicated that they have graded reading booklets (these were supplied to all intervention school teachers), nearly all teachers reported *using* the graded reading booklets (90 and 95 percent of teachers in the on-site and virtual arms respectively), and a high proportion reported using lesson plans provided by the government or a non-government organisation (81 and 89 percent of teachers in the on-site and virtual arms respectively). Moreover, Table A.12 shows that the print-richness of the classroom was substantially higher in both the on-site and virtual arms: a higher proportion of intervention classrooms had good quality posters and flashcards on the wall, and had storybooks.

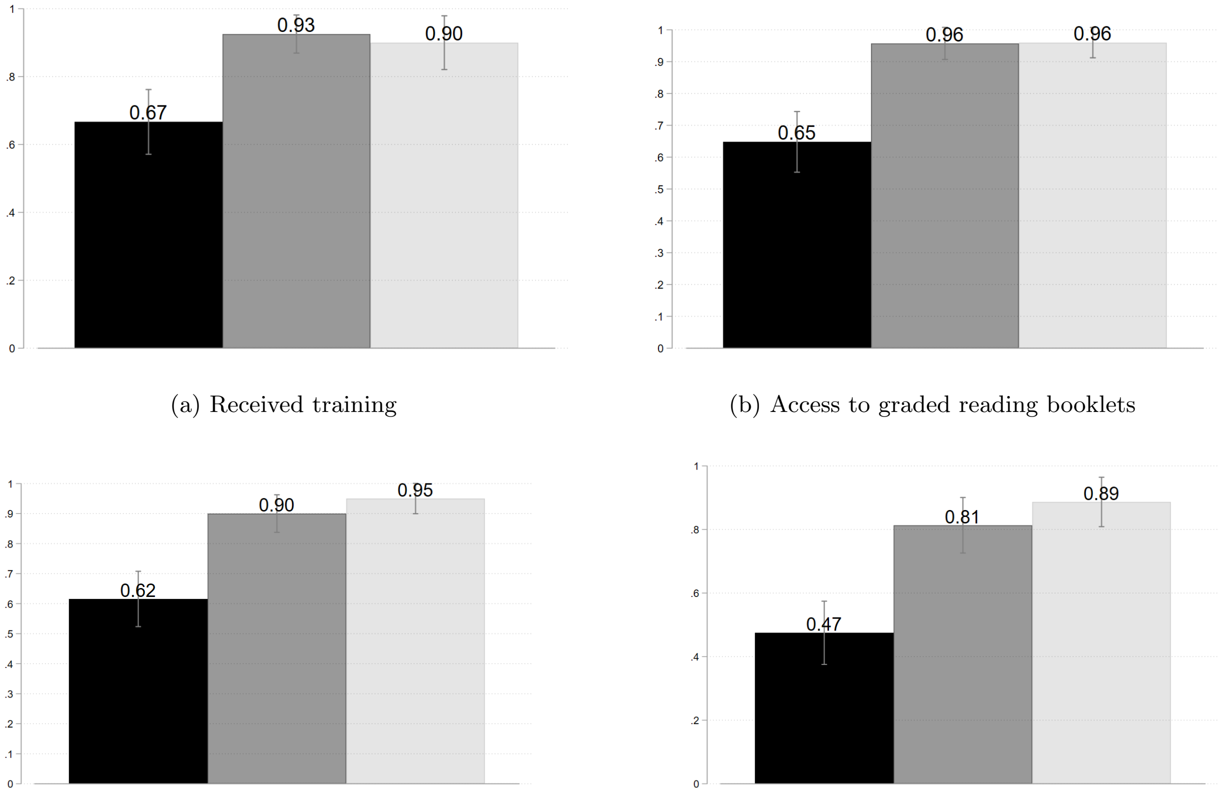
It is important to note that professional development activities were also taking place in the control schools. 67 percent of teachers in the control group report to have received training for ESL in 2019, which was most likely provided by the province or the district.[17] And 47 percent of control teachers also report 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, but 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 two main outcomes of interest: English oral language proficiency and English reading proficiency. Table 1, columns (1) and (4), show that the structured learning program with on-site coaching improved students' English oral language proficiency by 0.31 standard deviations, and improved reading proficiency by 0.13 standard deviations. These results are statistically significant at the one percent and the ten percent levels, respectively. The program therefore seems to have been

[17] 12 percent of the control school teachers responded that they received support from the National Education Collaboration Trust (NECT).The support by the NECT entails providing lesson plans which are very similar to the EGRS lesson plans and cascade training.

Figure 1: Quality of Implementation



(a) Received training

(b) Access to graded reading booklets

(c) Use graded reading booklets

(d) Use lesson plan

Note. Results from teacher questionnaire administered to 296 teachers in 180 schools. 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.

more effective at improving oral language proficiency than improving reading skills. 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. Table A.13 in the appendix shows that the magnitude of the effect sizes are slightly larger for the sample of students who are in grade three. This could either reflect the greater exposure to the program, or the fact that this sample of students is more responsive to the program overall. All subsequent analysis will be on the full sample of students.

Although it is encouraging that the virtual arm had an impact on oral language 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. Kotze et al. (2019) found that the virtual coaching program had a positive impact on students' oral language proficiency after the first year of the intervention. This suggests that the virtual coach was successful at facilitating teachers' adoption of teaching practices aimed to vocabulary development, but not decoding.

Table 1 further shows the results broken down by 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.[18] Nonetheless, performance in the on-site arm remains weak, with an average score of 25 percent in the comprehension test.

We perform three robustness checks. First, we show the treatment effects on the two main outcomes in Table A.14, including Inverse Probability Weights to compensate for any potential bias created through

[18] $0.06/0.19 = 0.31$

Table 1: Impact of on-site and virtual coaching on English oral and reading skills

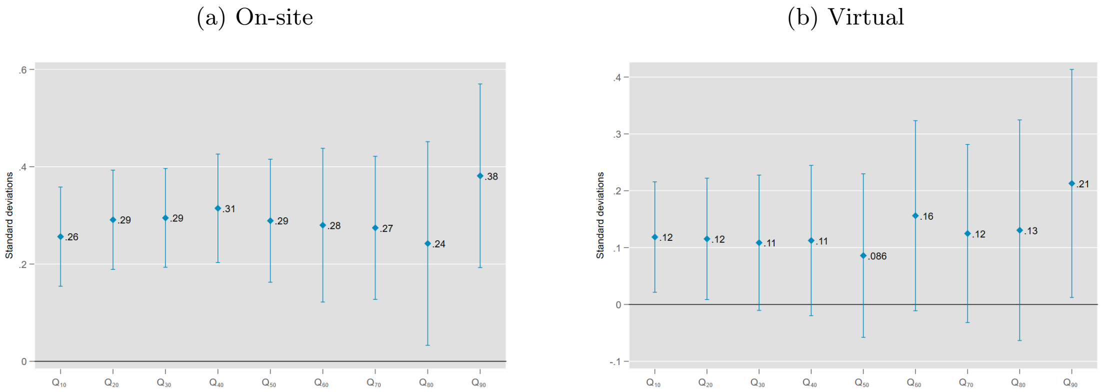
	Oral proficiency			Reading proficiency				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Index	Vocab	Listen compr.	Index	Word recog.	Read. fluency	Read. compr.	Written compr.
On-site coach	0.313***	0.420***	0.074***	0.130*	2.660*	2.458	0.060***	0.016
	(0.068)	(0.105)	(0.017)	(0.068)	(1.360)	(1.909)	(0.019)	(0.020)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

Notes: Each column is a separate regression, estimated using equation 1. Standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; 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 standardized to have control mean of zero and standard deviation of one. 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.

attrition. 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 show in Table A.15 that results are similar in magnitude and remain statistically significant when we do not weight each regression by the inverse of the number of students assessed at the end of year three. Third, Table A.16 shows that the treatment effects on English vocabulary do not depend on the choice of words used in the original instrument. As discussed in Section 4.1 we retested a subset of our sample using a more expanded set of words in the vocabulary test. Column (1) in Table A.16 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. It is is encouraging that there remains a positive treatment effect of similar magnitude (it is, in fact, slightly larger for both treatment arms) when using the expanded instrument.

Next, we explore the distribution of effect sizes, and find that the best-performing students benefited most from the program. Figures 2 and 3 report quantile regression results for each treatment on English oral and reading proficiency. Figure 2 shows that students across the distribution benefited in terms of oral language proficiency, but it is clear from figure 3, panel (a), that only the top half of students improved their English reading proficiency as a result of the on-site program. In fact, the effect sizes are monotonically

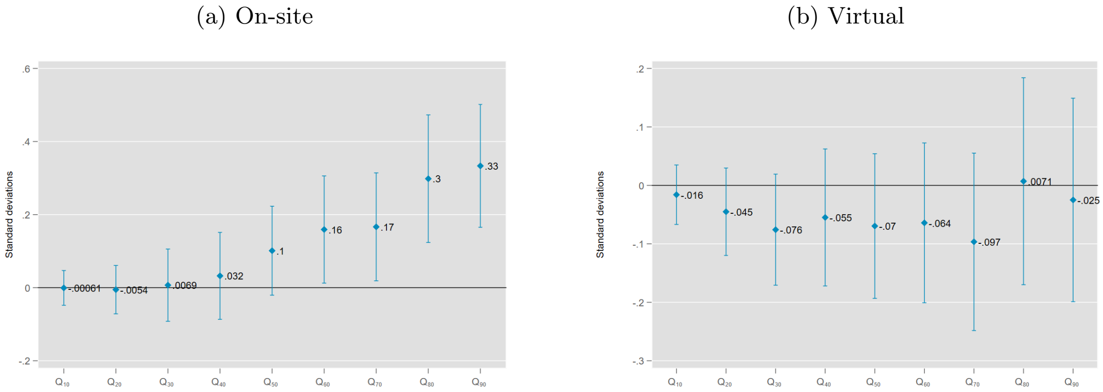
Figure 2: Quantile Regression— English Oral Proficiency


(a) On-site

(b) Virtual

Note. Estimates of quantile regressions, including same controls as equation (1), 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 3: Quantile Regression— English Reading Proficiency


(a) On-site

(b) Virtual

Note. See Figure 2.

increasing with each decile of reading proficiency. 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.

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. 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 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

Table 2: Home Language Literacy and Numeracy

	Home Language					Mathematics
	(1)	(2)	(3)	(4)	(5)	(6)
	Index	Letter Recognition	Oral Reading Fluency	Reading Comprehension	Written Comprehension	
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
Test:On-site=Virtual	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.

mathematics. We discuss possible reasons for this result in section 6.1 below.

5.3 Teaching knowledge and practice

Next, we investigate whether teacher knowledge and teaching practices changed as a result of the interventions. To summarize, results from both the teacher questionnaire and classroom observations data reveal that teachers in both treatment arms changed how frequently they implemented different activities in the classroom, and were more likely to provide individual feedback to students. However, only teachers that received on-site coaching were more likely to conduct group-guided reading— a teaching activity that is quite difficult to implement and facilitates learning by providing more opportunities for children to individually practice reading.

Table 3, column (1), shows that teachers in both treatment arms were more likely than the control teachers to correctly specify the number of times that they should repeat a phonics sound (the core methodologies specify three times). Columns (2)-(5) show that teachers were more likely to report that they implement the teaching activities at the required weekly frequency: phonics three times a week, group-guided reading daily, and writing four times a week, and shared reading twice a week. Note that this means that teachers in the treatment arms were more likely to conduct group-guided reading, but *less* likely to teach phonics: teachers in the control group were more than twice as likely to state that they teach phonics daily. The fact that the impact on group-guided reading is substantially higher in the on-site vis-a-vis the virtual arm, suggests that these self-reported data are unlikely to be exclusively driven by social desirability bias. But at the very least, these results are measures of knowledge: they show whether teachers know what they are

Table 3: Correct frequency of learning activities (self-reported)

	(1) Phonics sound	(2) Phonics lesson	(3) Group-guided reading	(4) Shared Reading	(5) Writing
On-site	0.351***	0.209***	0.469***	0.110	0.288***
	(0.073)	(0.072)	(0.067)	(0.067)	(0.071)
Virtual	0.255***	0.194**	0.240***	0.156**	0.280***
	(0.074)	(0.077)	(0.079)	(0.067)	(0.076)
Control mean	0.400	0.444	0.222	0.637	0.407
Observations	296	296	296	296	296
R-squared	0.141	0.084	0.172	0.066	0.092
Test:On-site=Virtual	0.239	0.861	0.007	0.516	0.913

Notes. Each column is a separate regression, estimated using equation 1. Data is at a teacher level, and standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects. The final row reports the p-value of the F-test of equality of coefficients

Table 4: Learning activities observed during the English lesson

	(1) Language Phonics	(2) Shared reading	(3) Paired reading	(4) Group-guided reading	(5) Pupil reads individually	(6) Writing	(7) Work individ. w/ students
On-site	-0.184*	-0.023	0.125	0.293*	0.333***	0.276**	0.268**
	(0.103)	(0.142)	(0.172)	(0.148)	(0.121)	(0.115)	(0.124)
Virtual	-0.179*	-0.296*	0.245	0.175	0.133	0.238*	0.257**
	(0.099)	(0.149)	(0.170)	(0.156)	(0.120)	(0.124)	(0.122)
Observations	53	53	53	53	53	53	47
R-squared	0.208	0.309	0.232	0.268	0.345	0.230	0.292
Control mean	1.000	0.778	0.278	0.167	0.056	0.722	0.000
Test:On-site=Virtual	0.975	0.104	0.412	0.442	0.157	0.492	0.944

Notes. See Table 3. Data comes from classroom observations conducted when teacher was teaching ESL. The first six outcomes are dummy variables equal to one if the respective teaching activities took place at least once during the full duration of the lecture. The final outcome is a dummy variable equal to one if the enumerator observed a teacher working individually with a student, at either the 15^{th} or 40^{th} minute in the lecture. This data was missing for 6 of the 53 classroom observations.

supposed to do in the classroom.

Table 4 shows that the above results are broadly confirmed in the classroom observations data. Teachers in both treatment arms were *less* likely to teach vocabulary or phonics, but *more* likely to have the students practice writing. Moreover, only teachers in the on-site arm were more likely to practice group-guided reading. As a result, children in the on-site arm were far more likely to get a chance to read out loud individually during the lesson. The final column in Table 4 shows teachers in both treatment arms were far more likely to be observed working individually with students. Note that not a single teacher in the control was observed working individually with students.

There is suggestive evidence that teachers in both arms applied some of their improved teaching practices

Table 5: Learning activities observed during the home language lesson

	(1) Language Phonics	(2) Shared reading	(3) Paired reading	(4) Group-guided reading	(5) Pupil reads individually	(6) Writing	(7) Work individ. w/ students
On-site	-0.294**	0.256	0.030	0.169	0.269**	0.299*	0.182
	(0.127)	(0.188)	(0.178)	(0.185)	(0.120)	(0.168)	(0.200)
Virtual	-0.056	0.005	0.290*	0.072	0.249*	0.262	0.136
	(0.113)	(0.180)	(0.149)	(0.187)	(0.133)	(0.159)	(0.204)
Observations	44	44	44	44	44	44	42
R-squared	0.395	0.252	0.283	0.187	0.346	0.176	0.209
Control mean	0.882	0.529	0.647	0.235	0.059	0.176	0.176
Test:On-site=Virtual	0.116	0.202	0.118	0.592	0.892	0.845	0.814

Notes. Data comes from classroom observations conducted when teacher was teaching the home language lesson, restricted to teachers who were teaching both English and Home Language on the day of classroom observations. See Table 4 for additional information.

to home language instruction as well as the English lesson. Table 5 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.

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.

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 projects, 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.[19]

It is unlikely that the crowding out of home language professional development explains the result. Table 6 shows that 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. But teachers in the on-site arm were less likely to receive training, relative to the control. 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 control relative to the intervention teachers. If anything, Table 5 suggests the opposite: 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 6 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 6 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 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 4 shows that teachers in

[19] 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.

Table 6: 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
Test:On-site=Virtual	0.340	0.159	0.674	0.478	0.287

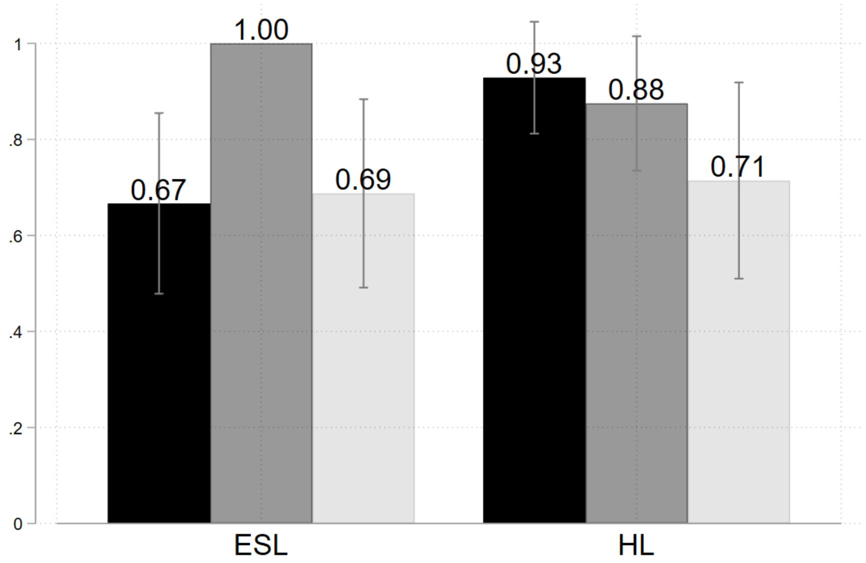
Notes. See Table 3. The outcome variables in the first three columns are dummy variables equal to one if a teacher (i) received professional development support in home language in 2019, (ii) uses HL lesson plans provided by an NGO, and (iii) has HL graded reading booklets, respectively. The outcome 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.

the virtual arm are less likely to be satisfied with how much they have progressed in the home language curriculum in the year of the study (71 percent vs 93 percent in the control), and all of these teachers refer to this program when explaining why they are struggling to complete the curriculum.[20] Given the small sample of teachers surveyed in the classroom observations who teach both ESL and HL the difference is not statistically significant, 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. It may be more appropriate to say that completing the curriculum requirement is a challenge within *actual* teaching time, rather than *allocated* teaching time, given that disruptions to teaching time are common in South Africa and were a reality at times during this project. Tablet usage analysis, which will be presented in the next section, indicate that teachers were increasingly less likely to spend time accessing those lessons scheduled for later in the term. This could plausibly be viewed as a proxy for curriculum coverage. If teachers struggle to complete the English lesson plans it is quite possible that this wold have led to some crowding out of Home Language teaching time. Although we do not have a similar proxy for curriculum coverage in the on-site coaching arm,

[20] 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.

Figure 4: 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.

it is likely that a similar challenge would have occurred. Why then do we not observe the same degree of crowding out of teaching time or a similar negative impact on Home language 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 Home Language lessons. We turn to this below.

6.2 Why was the virtual coaching intervention less effective?

Broadly defined, there are four possible explanations as to why the virtual arm was less effective: (i) the quality of implementation in the virtual arm was weaker; (ii) teachers interacted less frequently with the virtual coach, compared to the on-site coaches; (iii) the teachers faced barriers to accessing the technology; (iv) the virtual coach was unable to observe classroom practice thus weakening accountability, making it harder for a relationship of trust to form and harder to provide more targeted feedback. We demonstrate empirically that the first three explanations are unlikely to hold, and provide some 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 (Kotze et al., 2019). The teaching practices required to improve oral language proficiency —such as speaking to children in English and initiating songs and games— are likely to be easier than the teaching practices required to improve reading proficiency, such as group-guided reading. A likely interpretation is therefore that the virtual coach was effective at changing easier-to-implement teaching practice, but not the more difficult teaching practices.

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 less 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. And 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.

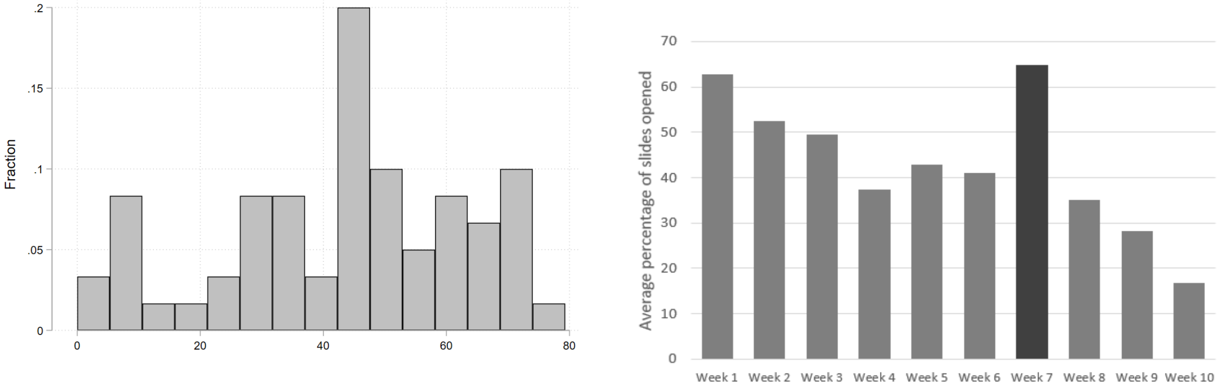
Third, analysis of tablet usage data suggests there were no barriers to accessing the technology: almost all teachers used the tablets, although at a variable rate. Panel (a) in Figure 5 shows a histogram of the distribution of percentage of term 3 lesson plan slides that were accessed by teachers any time between June and September 2019.[21] This might be considered a crude measure for potential curriculum coverage, or alternatively a proxy for intervention implementation fidelity. Evidently, there was quite a range of slide usage across teachers. The average teacher accessed 43 percent of the slides, only 10 percent of teachers accessed at least 70 percent of the slides, and only two teachers (3.3 percent) did not open a single slide during the third term.

A breakdown of slide coverage by each week of term 3 is even more revealing (panel (b), Figure 5). Interestingly, week seven was particularly well covered, and this happens to be the week in which assessments must take place.[22] It is also interesting that, aside from week 7, there seems to be a pattern of better coverage

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

[22] 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 5: Proportion of slides accessed on the tablet



(a) Histogram

(b) By week

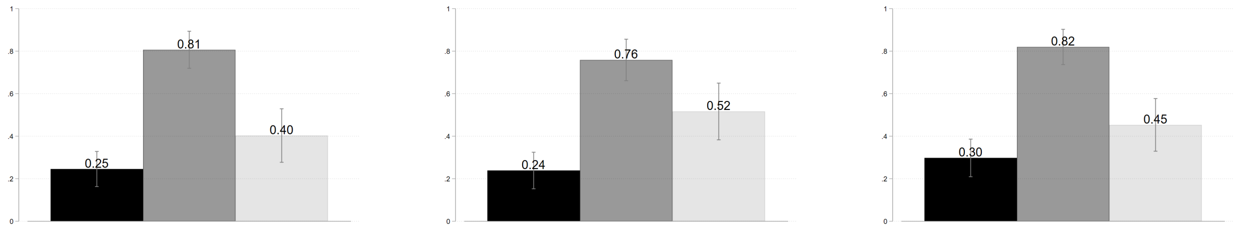
Note. Tablet usage data. The paper-based lesson plans were reformatted into pdf slides for teachers to navigate through on the tablet. Panel (a) shows a histogram of the proportion of slides that were opened by each teacher in the on-site arm, between July and September 2019. Panel (b) shows the proportion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

earlier in the term – weeks 1, 2 and 3 were the next best covered, with a steady decline in coverage until weeks 9 and 10 which had the lowest levels of coverage.

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

Results from the teacher surveys 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. Moreover, 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

Figure 6: Coaching support by teachers



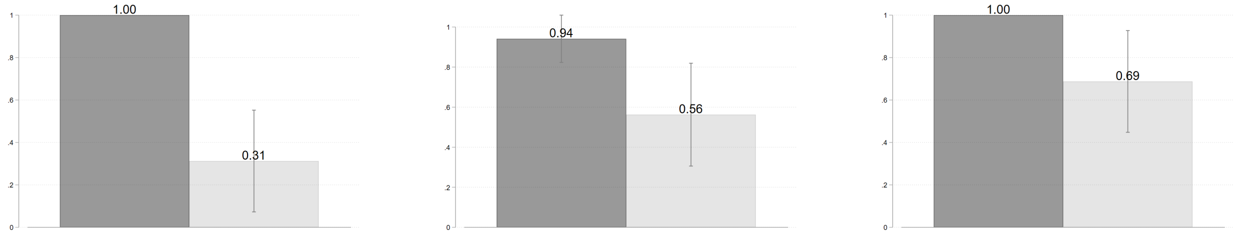
(a) Observed teaching

(b) Model lesson

(c) Received compliment

Note. See Figure 1

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

to mention the coach as someone who has helped her learn most this year.[23] Consistent with this result, teachers in the on-site arm were more satisfied with their curriculum coverage. Finally, the observed high variance in teacher participation in the virtual arm in voluntary activities —e.g. submissions to competitions, and reading the lesson plans— lends credence to the argument that the virtual coach was less able to hold teachers accountable, and the "opt-in" nature of most activities meant that only the more motivated teachers participated.

Finally, it is possible that the virtual arm was less effective because the coach was not able to observe teaching, thus limiting the ability to develop an accountable relationship of trust and the ability to provide targeted feedback. Teachers in the virtual arm were encouraged to submit videos to the coach, but very few did so consistently, and the videos did not cover the whole lesson. Qualitative fieldwork concluded that coaches in the on-site arm identified time management as a challenge in implementing the new lesson plans, and took steps to address that. This was not possible in the virtual arm. We therefore cannot rule out the possibility that the virtual arm would have been more effective if teachers were sufficiently incentivized to submit videos of their teaching on a regular basis. In fact, successful virtual coaching interventions evaluated

[23] Note that we did not specify in the pre-analysis plan that we will look at these outcomes, so this evidence should be treated as suggestive.

Table 7: Cost-effectiveness of the on-site and virtual coaching interventions

	On-site	**Virtual**
Costs per learner per year (USD)	66	51
Costs per teacher per year (USD)	2,747	2,131
Effect size on oral language proficiency per USD100	0.16	0.07
Effect size on reading proficiency per USD100	0.07	-

in the past all had a component of classroom observations (Allen et al., 2011; Bruns et al., 2017; Powell et al., 2010).[24]

6.3 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.[25] These estimates should therefore provide a realistic per-student cost if these models of delivery are scaled up. Based on these estimates, the per student costs of on-site coaching is USD66 per year and USD53 for virtual coaching. In terms of the cost of supporting a teacher per year, it is USD2,747 for on-site coaching and USD2,131 for virtual coaching.

Given the impacts of 0.31 on oral language proficiency and 0.13 on reading proficiency for on-site coaching over the three years of implementation, there was a 0.16 standard deviation increase in oral language proficiency for each USD100 spent and a 0.07 increase in reading proficiency. For virtual coaching, there was no significant impact on reading 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 more cost-effective than virtual coaching.

Figure B.8 in the appendix shows a more detailed breakdown of costs. The costs of the two interventions are roughly similar, since many components —such as training, program management costs, and teaching materials— are the same across the interventions. In fact, 49 percent of the costs in the on-site arm are costs that are also incurred in the virtual arm. The main differences between the two interventions are (i) the higher costs of salaries and transport for on-site coaching, (ii) the additional day of training for the virtual arm, and (iii) provision of tablets and hosting of software for the virtual arm. Note that these cost estimates suggest that a difference in average costs will not change with scale.

[24] For example, in a virtual coaching program for pre-kindergarten teachers implemented in 24 centers in the United States, virtual coaches met all the teachers in-person during training at the start of the program, and teachers were required to regularly submit videos of their teaching (Powell et al., 2010). And in a coaching program in Brazil, all teachers, both treatment and control, were observed once in the classroom at the beginning of the study and the treatment teachers received targeted feedback based on these observations (Bruns et al., 2017). And the "My Teaching Partner" in the United States also required teachers to send in videos of a full class of teaching twice a month (Allen et al., 2011).

[25] 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.

Table A.18 shows that the costs of the tablets (including maintenance and technical support) is 8.8 times higher than the cost of providing printed lesson plans. Tablets are often thought to be less expensive since they can be used for multiple years, but 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. Even if there were no maintenance/hosting costs, the tablets would only be more cost-effective in the long run if the lifetime of a tablet is four times longer than the lesson plans. Our costing data thus show that tablets are unlikely to be more cost-effective than paper-based lesson plans, although the tablets can provide other benefits, such as access to instructional videos.

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. Students in the on-site coaching treatment arm saw an effect size of 0.31 and 0.13 in oral language and reading proficiency skills respectively. This builds on an expanding body of evidence demonstrating the promise of structured pedagogy programs in improving early-grade reading proficiency, thus contributing to the external validity of these findings. Virtual coaching, however, had a much weaker effect on learning outcomes, with an impact of less than half of that of on-site coaching (0.12) for oral language proficiency and a negligible effect on reading proficiency. It also reduced home language reading proficiency, probably due to a crowding out of teaching time. We showed that the use of technology was not a barrier, but rather that a virtual coach was less effective in changing teacher instructional practice. We propose that is because a virtual coach faces greater barriers to developing a trusting relationship 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 countries like South Africa. This is not a convenient finding in contexts that have tight fiscal constraints or where re-prioritisation of public finances is difficult. However, in most education systems the wage bill accounts for upwards of 80 percent of education spending, and in these settings some degree of re-prioritization towards coaching is likely to improve the effectiveness of teachers, and in turn make overall education spending more cost-effective.

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. Our study shows that teachers in the virtual arm were held less accountable to complete their curriculum, relative to teachers in the virtual arm. The fact that teachers' usage of the lesson plans provided in the tablets almost doubled when they faced stronger incentives (the week when student assessments needed to be administered) suggests that the teachers in the virtual arm are still far from their production possibility frontier.

References

Allen, Joseph P, Robert C Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun, "An interaction-based approach to enhancing secondary school instruction and student achievement," *Science*, 2011, *333* (6045), 1034–1037.

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.

Evans, David K and Anna Popova, *What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews*, The World Bank, 2015.

Filmer, Deon, Margarita Langthaler, Robert Stehrer, and Thomas Vogel, "Learning to Realize Education's Promise," *World Development Report. The World Bank*, 2018.

Fleisch, Brahm, *The education triple cocktail: System-wide instructional reform in South Africa*, UCT Press/Juta and Company (Pty) Ltd, 2018.

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*, 2018, 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.

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, 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.

Appendix A Supplementary tables

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: Inter-rater reliability

	(1)	(2)	T-test	
	Original	Retest	Difference	Correlation
Variable	Mean/SE	Mean/SE	(1)-(2)	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.4: 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.5: Balance: School Characteristics

	(1)	(2)	(3)	(4)	T-test	
	Control	On-site	Virtual	Total	Difference	
Variable	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(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

	(1)	(2)	(3)	(4)	T-test	
	On-site	Control	Virtual	Total	Difference	
Variable	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(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

	(1)	(2)	(3)	(4)	T-test	
	Control	On-site	Virtual	Total	Difference	
Variable	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(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)		
No. letters sound correct	6.978	6.784	7.019	6.936	0.194	-0.041
	(0.447)	(0.590)	(0.610)	(0.307)		
Number of Words Read Correct	0.387	0.347	0.510	0.411	0.039	-0.123
	(0.096)	(0.103)	(0.148)	(0.066)		
Sentence Words Read 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

	(1)	(2)	(3)	(4)	T-test	
	Control	On-site	Virtual	Total	Difference	
Variable	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(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 Proficiency	-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: 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. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects.

Table A.10: Balance after attrition

	(1)	(2)	(3)	(4)	T-test	
	Control	On-site	Virtual	Total	Difference	
Variable	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(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)		
No. letters sound correct	6.983	7.006	7.101	7.023	-0.023	-0.118
	(0.442)	(0.633)	(0.632)	(0.315)		
Number of Words Read Correct	0.362	0.362	0.496	0.400	0.000	-0.134
	(0.093)	(0.116)	(0.150)	(0.067)		
Sentence Words Read 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.11: Probability reach 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. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects.

Table A.12: Print-richness of the classroom

	(1) Posters	(2) Flaschared	(3) Books (> 30)
On-site	0.215***	0.301***	0.316***
	(0.068)	(0.059)	(0.075)
Virtual	0.177**	0.157**	0.303***
	(0.068)	(0.072)	(0.081)
Control mean	0.617	0.609	0.256
Observations	292	292	292
R-squared	0.092	0.095	0.119
Test:On-site=Virtual	0.594	0.027	0.887

Notes. Standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects. For the first two columns the outcome variable is a binary variable equal to one if the enumerator coded that the quality of the ESL posters and flashcards are either of average or good quality. In the final column the outcome variable is binary indicating if the classroom has at least 30 reading books

Table A.13: Treatment effects on English oral and reading proficiency (grade 3 students only)

	Oral proficiency			Reading proficiency				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Index	Vocab	Listen compr.	Index	Word recog.	Read. fluency	Read. compr.	Written compr.
On-site coach	0.356***	0.468***	0.085***	0.179**	3.655***	3.465*	0.072***	0.031
	(0.078)	(0.115)	(0.020)	(0.072)	(1.397)	(2.013)	(0.023)	(0.022)
Virtual coach	0.149*	0.155	0.041**	-0.015	-0.648	0.466	0.028	-0.010
	(0.078)	(0.122)	(0.020)	(0.076)	(1.500)	(2.126)	(0.021)	(0.021)
Control mean	0.000	3.120	0.216	-0.000	23.121	27.255	0.191	0.355
Observations	2148	2148	2148	2109	2148	2148	2148	2109
R-squared	0.270	0.248	0.213	0.273	0.240	0.241	0.248	0.183
Test:On-site=Virtual	0.022	0.018	0.061	0.021	0.008	0.187	0.083	0.100

Notes: See table 1. Data restricted to students who reached grade 3 by the end of the third year.

Table A.14: Robustness Check: Using IPW weights

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

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: No student-level weights

	Oral proficiency			Reading proficiency				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Index	Vocab	Listen compr.	Index	Word recog.	Read. fluency	Read. compr.	Written compr.
On-site coach	0.288***	0.380***	0.069***	0.121*	2.483*	2.385	0.055***	0.013
	(0.065)	(0.099)	(0.017)	(0.069)	(1.313)	(1.873)	(0.018)	(0.021)
Virtual coach	0.118	0.126	0.032*	-0.054	-1.371	-0.927	0.012	-0.021
	(0.071)	(0.118)	(0.017)	(0.068)	(1.322)	(1.911)	(0.018)	(0.018)
Control mean	0.000	3.120	0.216	-0.000	23.121	27.255	0.191	0.355
Observations	2684	2684	2684	2632	2684	2684	2684	2632
R-squared	0.292	0.263	0.226	0.299	0.251	0.265	0.256	0.218
Test:On-site=Virtual	0.028	0.037	0.057	0.020	0.007	0.099	0.037	0.113

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

Table A.16: 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
Test:On-site=Virtual	0.085	0.329

Notes. Standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects.

Table A.17: Teacher behavior (home language classroom observations)

	(1) Instructions or questions	(2) Listen to reading	(3) Reading to students	(4) Writing on board	(5) Working with individual students
On-site	-0.087	-0.171	0.113	-0.313*	0.177
	(0.174)	(0.195)	(0.150)	(0.177)	(0.191)
Virtual	0.067	-0.295	0.199	-0.372**	0.117
	(0.179)	(0.174)	(0.146)	(0.172)	(0.196)
Observations	44	44	44	44	44
R-squared	0.326	0.236	0.179	0.326	0.216
Control mean	0.412	0.353	0.118	0.588	0.176
Test:On-site=Virtual	0.420	0.434	0.615	0.731	0.741

Notes. Standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects.

Table A.18: Comparing costs between paper-based and electronic lesson plans (USD)

Paper-based		**Electronic**	
Printing	3,638	Tablets	14,291
		Hosting and software maintenance	9,013
		Technical support	8,829
Total	3,638	Total	32,132

Notes. Costs are per year for about 86 teachers in the on-site treatment arm and 82 teachers in the virtual treatment arm. All costs denoted in USD.

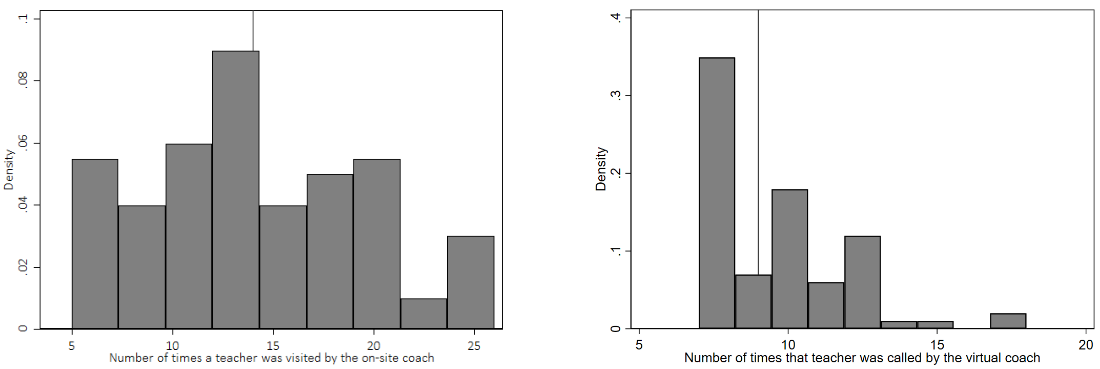
Table A.19: Comparing costs between on-site and virtual coaching (USD)

	on-site	**virtual**
Coach salaries	96,529	24,980
Travel and accommodation	31,491	1,947
Coach communication		1,735
Teacher data		9,322
Total	128,019	37,985

Notes. See Table A.18.

Appendix B Supplementary figures

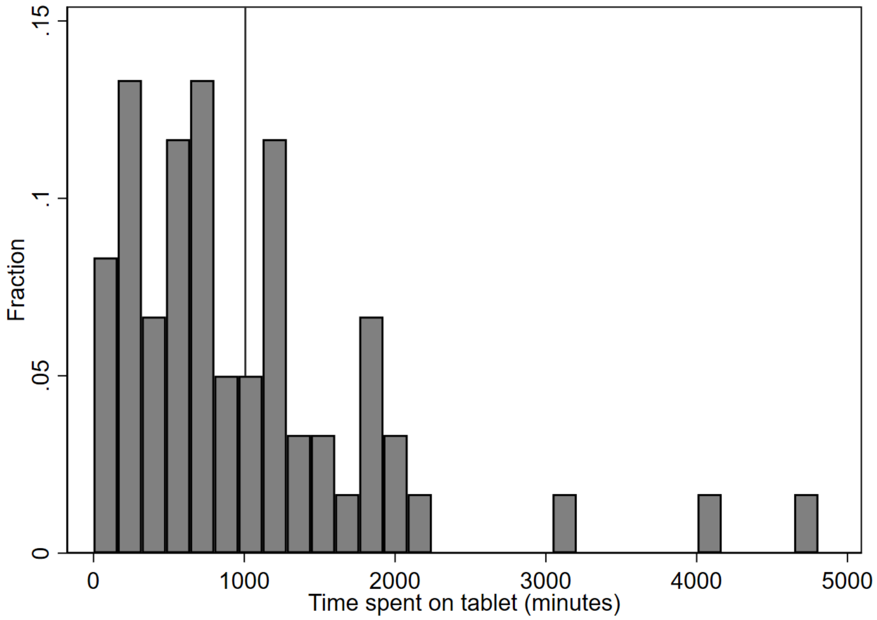
Figure B.1: Histogram of number of times that a teacher was visited or called by a coach



(a) on-site

(b) virtual

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

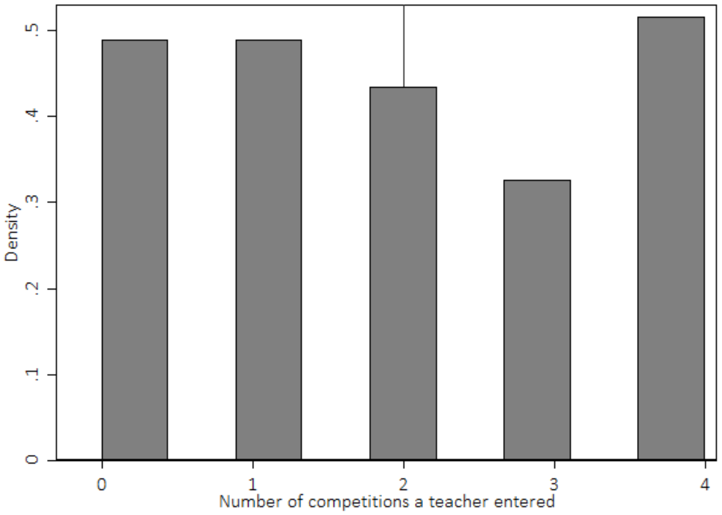
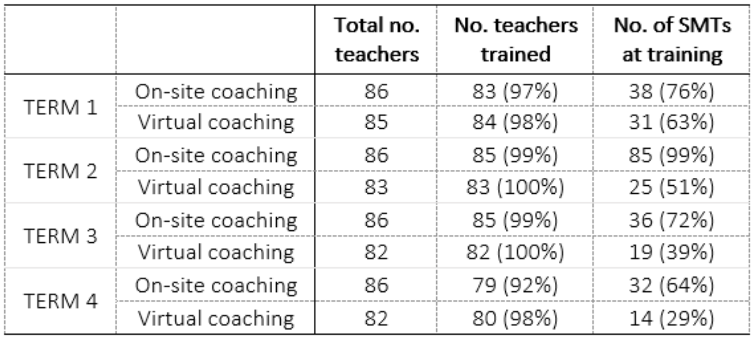
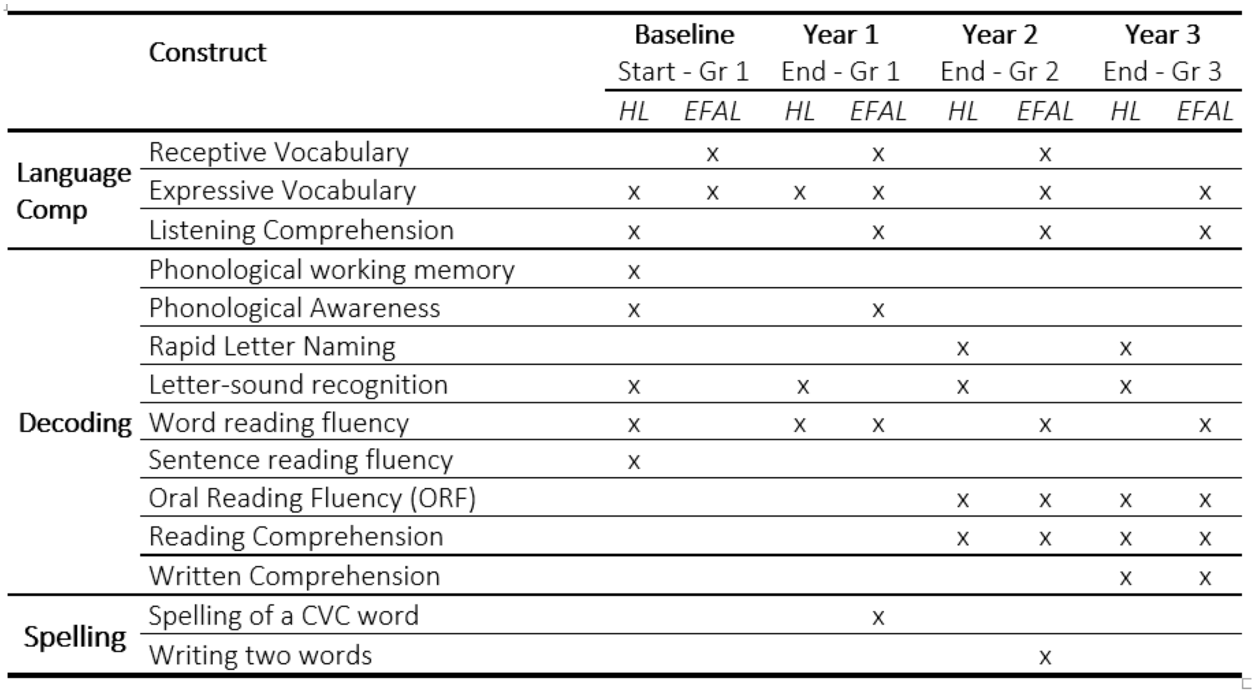


Figure B.4: Summary of attendance at training sessions



		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%)

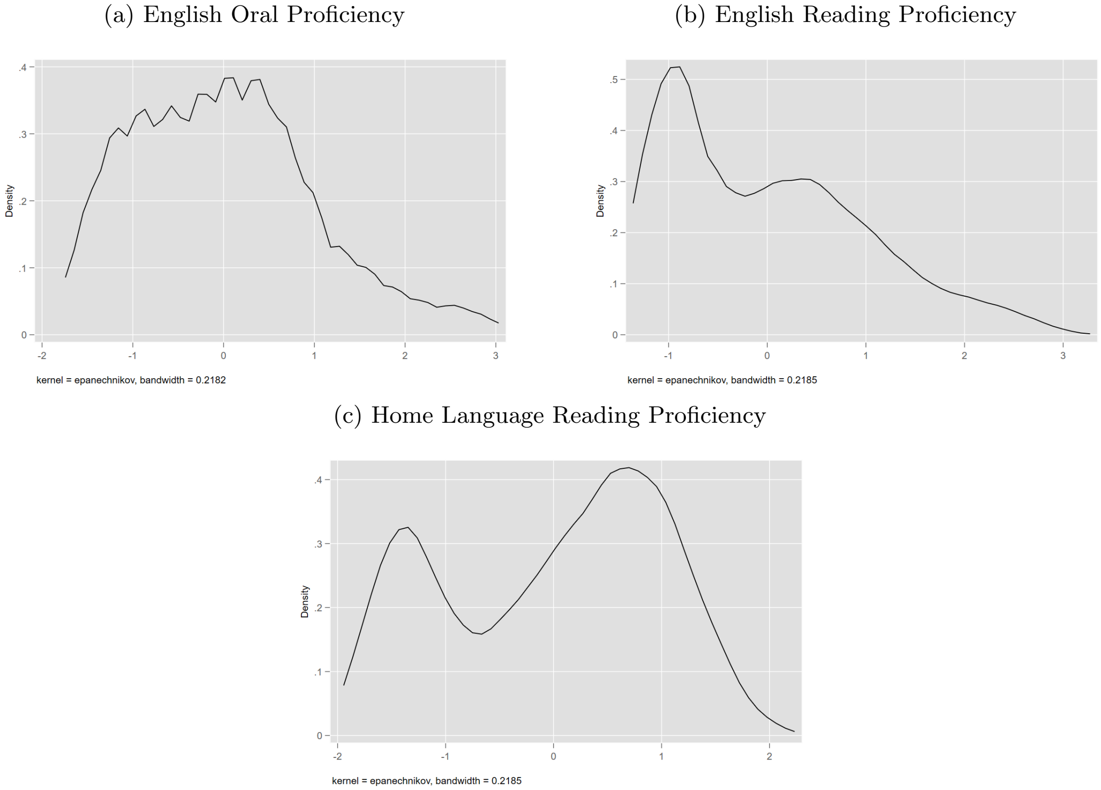
Figure B.5: Summary of student assessment, by domain, grade, and subject



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

Note. HL = Home Language; EFAL = English as a First Additional Language.

Figure B.6: 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. 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 letter recognition, oral reading fluency, reading comprehension, and written comprehension subtasks. Data is restricted to the control group.

Figure B.7: Percentage of students in the control group who could not read a single word

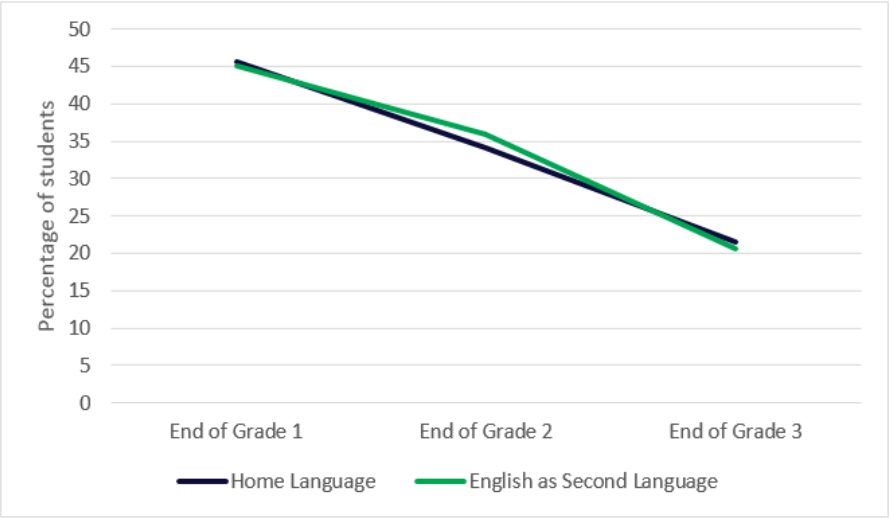
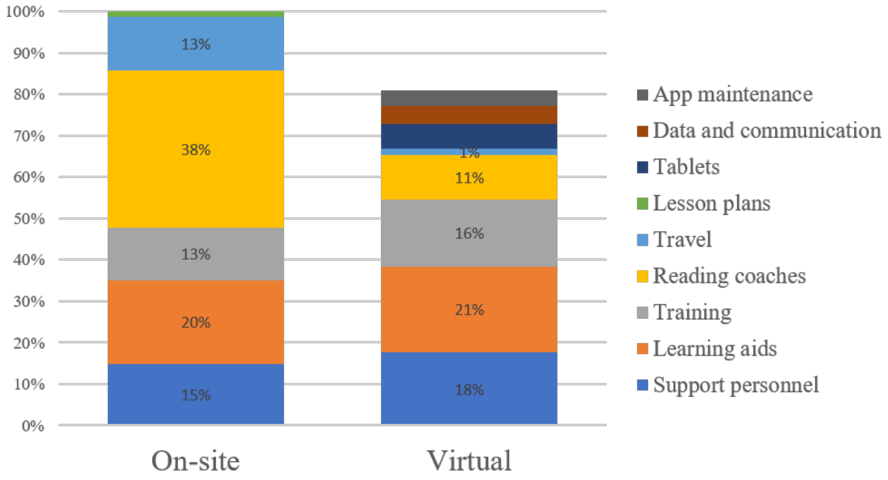


Figure B.8: Breakdown of cost drivers



Note. Costs as a proportion of total cost in the on-site arm.