

The sustainability of early education interventions: Do learning gains and improved teacher practices persist?

Jacobus Cilliers, Brahm Fleisch, Janeli Kotze, Mpumi
Mohohlwane, and Stephen Taylor

May
2019

Abstract

The cost-effectiveness of one-time investments in teacher productivity relies crucially on two different sources of persistence: the future life-time benefits to students exposed to the teachers, and the sustained productivity of teachers. This study tests for both these using results from a randomized evaluation of two different forms of teacher professional development - Coaching or Training - aimed at improving teaching of early-grade reading in South Africa. We find that for both programs the initial improvements in learning persist, when students are assessed one year after the program ended. However, the impacts on the subsequent cohort of students – i.e. those who were taught by teachers one year after they received coaching/training – is roughly half the size and is only statistically significant for schools where teachers received Coaching. In both programs improvements in teacher knowledge and use of resources persisted up to three years after participating in the program, but only teachers who received Coaching maintain their improved teaching practices. These results demonstrate that resource-intensive programs that induce sustained changes in productivity can be more cost-effective in the long run due to benefits for future cohorts.

Introduction

Investing in student learning at early grades is often motivated by the idea that early gains can compound over time. Core skills such as reading and arithmetic are hierarchical, such that one must master the basics before they can expect to learn more advanced skills (Heckman, 2006). Moreover, the ability to read with comprehension enables learning in other domains. However, it is also theoretically possible that early gains could remain constant or even fade over time, if learning requires dynamic, complementary investments to consolidate the knowledge (Heckman, 2007). There could also be learning plateaus and students who were not exposed to the intervention may eventually catch up to their peers who received added early investments.

An equally important question of program sustainability is that of the *producers* of learning themselves: the teachers. Investments in the quality of teaching have the potential to benefit subsequent cohorts of students, if improved teaching practices are sustained after the program ends. Conclusions around the cost-effectiveness - and thus scalability - of different interventions clearly depend on whether a one-off investment can induce a lasting change in productivity, thus benefiting future generations. A relatively more expensive intervention could become more cost-effective when considering the flow of future benefits. Or low-cost investments in future years can be highly cost-effective if they help sustain the initial gains in productivity. This insight is well-understood when considering investments in physical capital, but less so with human capital. Academics and policy-makers often make conclusions based on evidence for only the first cohort of beneficiaries.

There is surprisingly little evidence on either of these two dimensions of sustainability. A deworming program in Kenya has been shown to have improved both educational attainment and labor market performance in later years (Baird et al, 2016). High-quality, early childhood development and pre-school programs have also been found to impact adult labor market outcomes (e.g. Heckman et al, 2010; Gertler et al, 2014), although the initial gains in learning faded after a few years. In South Africa, a historical policy change leading to more years of mother-tongue instruction in primary school was found to have improved labor market outcomes for those affected (Eriksson, 2014). However, a review of educational impact evaluations by McEwan (2015) suggests that only 10 percent of studies report effects measured more than one month after program implementation. It is also true that meta-analyses and systematic reviews of educational interventions typically compare effect sizes at the time of program completion. More importantly, there is little credible evidence on the persistence of interventions that target teachers at early grades.

This paper reports on these two important and distinct aspects of sustainability through a longitudinal analysis of the teachers and students who participated in a randomized evaluation of two programs aimed at improving teaching of home language (Setswana) early-grade literacy in 180 public primary schools South Africa. Of these 180 schools, 50 were randomly assigned to receive residential teacher Training workshops (two days at a time, twice a year), 50 were assigned to receive on-site specialist Coaching to teachers, and 100 schools were assigned to the control. In both programs, teachers received the same lesson plans and learning aids. A different grade of teachers participated every year (grade one teachers in the first year, grade two and three teachers in the second and third years respectively), so most teachers were only directly exposed to the program for one year. After two years of implementation (at the end of 2016) Training and Coaching had a positively impact on learning of 0.12 and 0.24 standard deviations (SD) respectively (Cilliers et al, 2018). Moreover, lesson observations and document inspection revealed significant improvements in teacher practice, especially among instructional methods which teachers generally find harder to implement.

Towards the end of 2018, roughly a year after the completion of the program, a follow-up data collection was conducted to investigate the impacts on the original sample of learners – the first cohort – who had been tracked since they started grade one in 2015. The survey also tested a fresh sample of grade three children in the same schools – the second cohort – and conducted interviews with grade three teachers, in order to see if teacher knowledge and reported practices were significantly different compared to the control group one year after program support had ended.

There are three main results to highlight in this paper. First, the impacts on the first cohort of learners persisted one year after the programs ended (when most students were in grade 4), with both the Training and Coaching groups of children still performing better on Home Language literacy than those in the control group. The effect size for the Coaching group (0.25 SD) is almost exactly the same as that observed directly after two years of intervention (0.24 SD). The effect size for the Training group increased slightly from 0.12 SD to 0.20 SD, although we cannot reject the null hypothesis that the two coefficients are the same. This result means there is neither clear evidence of a substantial compounding of early learning benefits nor evidence of an obvious fadeout of effects. However, future tracking of this panel of students will inform conclusions about the long run benefits of an early boost to home language reading on later educational outcomes. Although the relative effect sizes for students in the two treatment groups did not change dramatically, they have changed enough to illustrate the potential impact that measuring persistence of effects can have on cost-effectiveness estimates of such programs. Whereas in

2016 the Coaching intervention was relatively more cost-effective, by 2018 the gap between Training and Coaching appears to have narrowed somewhat placing the two interventions in a similar range of cost-effectiveness.

Second, both programs had positive impacts on learning for the second cohort of grade three students – 0.11 and 0.17 SD for the Teaching and Coaching arms respectively – although the impact was only statistically significant for Coaching. A one-year investment in teacher productivity, through regular visits from reading coaches, had a persistent positive impact on student learning for at least one year after the termination of the program. Moreover, this might be an underestimate of the true impact of the programs, due to teacher turnover – only 77 percent of our sample of grade three teachers in the treatment schools participated in the program the previous year – and the fact that treated teachers were less likely to receive teacher professional development the following year. Our results should be interpreted as an Intent-to-Treat estimate.

Third, the use of resources provided by the program, and improved knowledge and teaching practice persisted after the program ended, especially for those teachers who had received ongoing in-classroom coaching. This result is consistent with, and provides insight into, the observed effects on learner outcomes mentioned above. Teachers in the Coaching and Training schools were still using the resources provided by the program – the lesson plans and reading materials – and scored better than the control teachers on a simple pedagogical test. Although both Training and Coaching had a positive impact on the proportion of teachers who implement different instructional methods correctly, this effect fades out after one year for teachers that received Training, but not for Coaching. This exact same trend is observed for the print-richness of the classroom environment and the number of and difficulty of exercises completed by students. Moreover, even grade one teachers, who were exposed to the program three years ago, demonstrate improved knowledge and use of materials. These results suggest that in order to achieve sustained and meaningful change in teacher practices that also lead to better student learning outcomes, teachers require the ongoing in-classroom support offered by coaches.

We also find evidence of positive spillovers in teacher knowledge and use of resources. Teachers in treated schools who did not teach grade three the previous year (and were hence not exposed to the program) demonstrate better pedagogical knowledge and use of resources, compared to similar teachers in the control. Almost all of them use the lesson plans provided by the program, but there is no evidence that

they behave differently inside the classroom. It is thus possible that the program has benefits to the school as a whole through the resources provided.¹

These results extend the existing literature on both the sustainability of the impacts of educational interventions and understanding how this is affected by the mechanics of changing teacher practice. One important evaluation of a national Coaching program in Peru found that the impact of coaching persisted even when teachers moved schools, but that the benefits are lost to the school that loses the teacher (Majerowicz and Montero, 2018). Our study extends these findings by measuring the impacts on both learning outcomes and teaching practice for teachers who had received coaching one year earlier (the grade three cohort of 2018) and those coached three years earlier (the grade one cohort of 2018). Moreover, we compare all these matters of sustainability for one commonly used modality of professional development (centralized training workshops) with another increasingly favored modality (on-site coaching).

Program description and evaluation design

This paper is based on the evaluation of two different modes of teacher professional development conducted in a sample of 180 public primary schools that serve poor communities in South Africa's North West province. In both programs, teachers received the same lesson plans and supporting materials, such as graded readers, flash cards, and posters. The lesson plans cover both the content of the curriculum and the prescribed pedagogical practices.² But the programs differed in the modality of delivery. In the Training arm, teachers received training twice per year on how to teach early-grade reading in one's home language. In the Coaching arm, teachers received monthly visits from reading coaches who observed teaching, provided feedback on how to improve pedagogical techniques, and demonstrated correct teaching techniques. We randomly allocated 50 schools to participate in each treatment program and 80 schools to the control. Cilliers et al (2019) provides more details of the interventions and evaluation design.

The programs were implemented over a period of three years. Each year teachers of a different grade participated in the program: The grade one teachers received Training/Coaching in the first year, the grade two teachers participated in the second year, and grade three teachers in the third year. Our first cohort of students were therefore always taught by teachers who were receiving either Training or Coaching that

¹ Note that positive spillovers mean we cannot use an instrument variable approach to construct the local average treatment effect for teachers exposed to the program.

² The curriculum in South Africa is highly detailed and prescriptive, specifying the weekly frequency and time that should be dedicated to different teaching activities.

same year. Whereas our first cohort was exposed to the program over three years, the teachers themselves were normally only exposed to one year, since a different teacher typically teaches in a different grade.

Conceptual framework

Donor programs are typically once-off investments in human and physical capital that improve the production efficiency of a good or service. Examples include both infrastructure investments, such as building schools or roads, and training of service providers. The intention is that the benefits will persist over time, subject to some rate of depreciation. When doing cost-effectiveness or cost-benefit analysis of a one-time investment, one clearly needs to consider the future stream of benefits. An expensive program might become the most cost-effective, relative to a less-expensive program, only when evaluated over multiple years, because its net benefits depreciate at a lower rate. This insight is understood in case of infrastructure programs, but it is rarely explicitly considered when it comes to improving human capital. Many evaluations that look at long-term impacts of programs, for example, assess the sustained impacts on the same cohort of beneficiaries, rather than on the impact on future beneficiaries. There is a risk that donors are overly myopic, given the time frame over which their programs are evaluated, and their performance is assessed.

This study considers both sources of persistence – on both the producers’ and beneficiaries’ side – in the context of teacher professional development in teaching of early-grade home-language literacy.

Persistence in teacher productivity

Both programs in this study could improve learning through a combination of theoretical channels: additional resources, improved teacher knowledge, higher teacher effort, or new teaching skills. Each of these channels imply different possible rates of depreciation. For example, if the impacts are purely driven by the additional resources that the teachers from both interventions received – the lesson plans and graded readers – then one would expect a relatively high level of persistence, provided that teachers require no additional skills or effort to appropriately utilize these resources, and the resources do not deteriorate quickly. Similarly, if the impacts were driven purely by improved content knowledge, then one would expect high levels of persistence, provided that teachers do not forget what they learnt.

One might expect weak persistence if the program induced teachers to initially increase their effort level – such as applying more difficult teaching techniques in the classroom, or spending more time preparing for class – but there is no mechanism to permanently shift the *costs* of or *returns* to effort. For example, if the Coaching changed behavior, only because the monitoring induced higher levels of effort, then one would expect only transient behavior changes. However, there could be a more permanent change if the

coaching helped teachers develop new habits and improve their skills, which would reduce the effort cost of implementing the techniques. The returns to effort could also increase when teachers observe the improvements in student learning due to their efforts.

It is thus possible that these two programs have very different dynamic impacts on productivity, depending on the mechanisms at play.

Persistence in learning

The potential trends in learning for students who were exposed to an early-grade reading intervention is theoretically ambiguous. Pfof et al (2014) posit three hypotheses associated the relationship between early reading performance and later reading achievement. The “Matthew effect” or fan-spread hypothesis (Figure 1) assumes a positive relationship between successful early reading achievement and later reading proficiency. Learners that read better in the first two grades continue to show higher achievement later in their school careers due to the mechanisms of bootstrapping and the reciprocal relationship between better reading and greater learning across the curriculum. A second hypothesis associated with relationship between initial reading performances and later reading levels suggests that the relationship would be negative, sometimes referred to as the “developmental lag” model (Figure 2). Learners who perform well in the early grades continue to improve, but at a lower rate than initially poorly performing learners. This would be the case if there is a learning threshold that is eventually reached by all learners. An early intervention merely speeds up the pace at which learners reach the threshold. This produces the opposite of the Matthew effect. The third hypothesis suggests a stable proficiency difference between high and low performing learners.

Although the three hypotheses are specifically framed to explore the patterns of achievement of a learner population over time, the three hypotheses equally apply to tracking the sustainability of the impact of early grade reading interventions over time. If the early grade intervention is initially successful, then the learner population sample that have benefited should make greater gains overtime. With the second hypothesis, while the intervention benefited learners in the short-term, the learners that had not been part of the intervention would ‘catch-up’ over time, reducing the learning gaps between those that had originally benefited from the intervention and those that had not. The third hypothesis assumes a stable pattern of performance between those that initially benefitted and those who did not. This study provides empirical evidence to contribute to advancing knowledge on this critical question.

Data and Methodology

Data collection

There were four rounds of data collection: the first prior to the start of the interventions (February 2015), the second was at the end of the first year of implementation (November 2015), the third was at the end of the second year of implementation (November 2016), and the fourth was a year after the interventions had concluded (August 2018). For the first cohort, we randomly sampled 20 students per school and tested them at the start of grade one. We tracked and reassessed the same students at the end of grade one, the end of grade two, and again in the third term of grade four. Of the 3,539 students sampled at baseline, we reassessed 2,590 students in grade four. During the fourth round of data collection, we also randomly sampled and assessed a second cohort of grade three students, 10 per school. We only assessed these students in a reduced set of 164 school, yielding a sample of 1,636 grade three students.³

During the school visits we administered student assessments, a principal questionnaire, and teacher questionnaires with the grade one and grade three teachers. The student assessments for grade three and grade four were mostly similar and included an EGRA-type orally administered assessment, as well as a written assessment administered in a group setting. The orally administered assessment included letter recognition, word reading, and oral reading fluency tasks, all based on the Setswana EGRA instrument, whereas the written assessment included a reading comprehension task taken from the 2015 PIRLS Literacy assessment. The letter recognition, word recognition, and oral reading fluency tasks are very similar to the assessments we used during the third round of data collection (testing the learners at the end of grade two), however the oral reading fluency passage used in the different grades were selected to be grade appropriate. Two tasks, phonological awareness and non-word recognition, were only assessed at grades one and two, since they are less appropriate for later grade. For each round of assessment, we combined all the sub-tasks into one aggregate score using principal components. We then created a Z-score by subtracting the mean and standard deviation in the control. Results on the mean index can therefore be interpreted in terms of standard deviation increases.

The teacher questionnaire included a basic pedagogical knowledge question: how often should one repeat a word when teaching phonics? It also asked teachers to report the weekly frequency with which they implement the following activities: group-guided reading, phonics, spelling, and writing. The official curriculum prescribes the exact frequency that these activities should take place at each grade, which is

³ 16 of our schools are multi-grade schools, and the Coaching and Training interventions were not designed to work in these settings. We are in the process of evaluating a scaled-up version of the program and decided to restrict all future data collection to these schools. Data capture error means that for some schools we did not receive data on 10 pupils.

reinforced by the training, coaching and lesson plans. Our variable labeled “teaching activities” is the proportion of activities that a teacher implements at the appropriate frequency.

In addition to the teacher questionnaire, fieldworkers also observed the classroom and inspected documents. Fieldworkers were required to rate on a four-point scale the availability of the following resources in the classroom: a reading corner, story books, Setswana posters, picture posters, and flash cards. We use these variables to construct a Kling index of the print-richness of the classroom. Fieldworkers also asked the teacher to provide the exercise book and government-issued handbook of one of the best-performing students in the class.⁴ They then counted the number of days that the pupil completed any exercise, completed a writing exercise, wrote a full sentence, and wrote a full paragraph. They also counted the total number of pages completed in the handbook. We construct a Kling index of all these indicators.

While we tracked the first cohort of students over four years, we interviewed a different set of teachers each year, because students typically have different teachers in different grades. In the first three rounds of data collection, we also surveyed the teachers of the first cohort. However, in the final round of data collection we decided to only survey grade one and grade three teachers, even though our first cohort were in grade four. The motivation was to look at sustainability: the grade three teachers would have been exposed to the intervention the *prior* year, provided that the teachers are still teaching the same grade at the same school.

Empirical strategy

We estimate the impact on each cohort separate, using the following equation:

$$y_{icsb1} = \beta_0 + \beta_1 (\text{Training})_s + \beta_2 (\text{Coaching})_s + X'_{ib0} \Gamma + \rho_b + \epsilon_{icsb1}, \quad (1)$$

where y_{icsb1} is the aggregate score of reading proficiency for student i who was taught by a teacher in class c , school s and stratum b ; $(\text{Training})_s$ and $(\text{Coaching})_s$ are the relevant treatment dummies; ρ_b refers to strata fixed effects; X_{ib0} is a vector of controls; and ϵ_{icsb1} is the error term clustered at the school-level. When estimating the impact on the first cohort, we include the same set of controls as used by Cilliers et al (2019). For the second cohort we do not have baseline data, and so only control for student gender and age, and the same set of community characteristics. In both specifications we demean all our controls and interact with the treatment dummies, as proposed by Gibbons et al (2018), Lin (2013), and Imben and Rubin (2015).

When examining teacher-level effects, we estimate the following model:

⁴ We asked for the written material of the one of the best-performing students in order to limit the risk of bias: the risk that teachers in the treatment arms are more likely to share the exercise of their best students.

$$M_{cs} = \alpha_1 + \beta_1(\text{Training})_s + \beta_2(\text{Coaching})_s + \rho_b + \epsilon_{csb1}, \quad (3)$$

where M_{cs} is the mediating variable of interest for a teacher in class c and school s . Standard errors are clustered at the school-level.

Attrition and balance

Figure 3 shows both grade progression and attrition by treatment arm for the first cohort. Roughly 28 percent of the sample have attrited and roughly a fifth of the sample have repeated a grade. This does not vary substantially by treatment status. Column (1) in Table 1 confirms that there is no statistically significant difference in the rate of attrition across treatment arms. Columns (2) and (3) show that older students and boys are more likely to attrite. This is unsurprising since older students and males perform worse on average and are therefore more likely to drop out. However, the coefficients on the interaction terms show that both treatment arms off-set this trend: i.e. older students and boys are *less* likely to attrite in the treatment arms, relative to the control. This is an encouraging result, although it creates some concerns for the comparability of the sample.⁵ In all analysis we control for both age and gender, interacted with treatment. Column (4) shows that the attrition rate does not vary systematically by baseline reading proficiency.

Table 2, columns (1) to (3), reports balance on the sub-set of students that remain in the sample by end of grade 4.⁶ There is slight imbalance in the Training arm. Since the Training arm is more likely to have boys and older students – typically under-performers in South African schools – relative to the control, this could bias the result downwards. But it is encouraging that the sample is not imbalanced on baseline reading proficiency. To offset possible bias from imbalance, we control for all these variables, interacted with treatment dummies, in our main specification. Since we do not have baseline data for the later cohort, we only report balance on two time-invariant student characteristics: age and gender. Table 2, columns (4) and (5), shows that students in the Training arm are slightly older, but the sample is balanced otherwise. Again, we control for both age and gender in all our analysis.

⁵ Note that this would most likely provide an under-estimate of the result, since the remaining sample in the treatment arm are weaker students on average.

⁶ We refer the readers to Cilliers et al (2019) for balanced statistics on a larger set of variables.

Results

First cohort

Table 3 shows the results on reading proficiency for the first cohort of students, evaluated at the end of grade two and end of grade four respectively.⁷ To ensure comparability, we restrict the sample to the same students who were assessed in both these rounds of data collection (results do not change in any meaningful way with the full samples). The impact of Coaching is surprisingly constant between grade two and grade four: 0.246 and 0.25 standard deviations, respectively. In contrast, the magnitude of effect size in the Training arm has increased by 37 percent - from 0.145 to 0.2 standard deviations. Students taught by teachers who received Training have thus caught up almost entirely to students in the Coaching arm. Columns (3) to (11) show results for each reading task. The largest gain in the Training is in the higher-order domains of reading proficiency: paragraph reading and reading comprehension. These are the areas where students in the Training arm lagged the furthest behind students in the Coaching arm at end of grade two. Coaching still substantially out-performs Training in written comprehension. This assessment was only conducted end of grade four, so we cannot compare it to results in grade two.

Figure 4 shows the predicted number of words read in a paragraph, for each year and each evaluation arm for our first cohort based on the coefficient estimates from our main specification but pooled over all the years of data collection: grades one, two, and four. We choose paragraph reading, since this metric is roughly comparable across years, and arguably the best indicator for reading fluency. Students in the Coaching arm learned at a faster pace compared to Training in the first two years of the study, but students in the Training arm learned at a faster pace compared to Coaching in the second two years. The net result is that the magnitude of the treatment effect is only slightly smaller for Training relative to Coaching: an additional 7.8 versus 7.4 words read. These effect sizes surmount to 17% and 16% of the learning that took place in the control over those four years.⁸

We therefore do not find any evidence of diminishing impacts, nor of compounding. We cannot draw strong interpretations on the dynamics of learning *after* the program ended, since we do not know what the magnitudes of treatment effects were at the end of grade three – the final year of exposure to the program. The relative catch-up of students in the Training arm, for example, could have happened in the final year of exposure to the intervention. Most likely there is some fade-out of results in the Coaching arm, since it is reasonable to assume that the intervention also had a positive impact on learning in the

⁷ We were unable to assess them the end of grade 3.

⁸ The average number of words read in the control at the start of grade 1 was 2. The average at the end of grade 4 was 47, implying an increase of 45 words.

third year. Nonetheless, the fact that the magnitude of effect sizes remains high after the end of grade 4 is encouraging for the expected long-term impacts for the students exposed to the program.

Second cohort

Table 4 shows results for the second cohort of students, who were assessed at the end of grade three. These students were taught by teachers who were likely to have received the intervention in the *previous* year. If the improved teaching practices persist in subsequent years this should be reflected by the learning gains for these students as well. Since we did not assess the second cohort at the beginning of grade one, we cannot control for baseline learning. As a result, we have far less statistical power to detect statistically significant impacts. Nonetheless, strong patterns emerge.

Students in the Coaching arm learned on average 0.167 standard deviations more compared to the control. In contrast, the effect size of Training is 0.11 and this is not statistically significant at conventional levels of significance. For both treatment arms the strongest gains are in reading comprehension – arguably the most important indicator of reading proficiency. It is encouraging that Coaching had an impact on the subsequent cohort of students, but there remains substantial fade-out of impacts. Recall that we only assessed grade three students in a subset of 164 out of 180 schools. As a rough comparison, Table A.1 in the appendix shows the impacts for first cohort evaluated at the end of grade two, but estimated on the same sample of schools. The effect size of Coaching was 0.35 standard deviations. The impact therefore roughly *halved* in the second year.

Cost-effectiveness

Although Cilliers et al (2019) found that coaching was the more cost-effective intervention when evaluating students at the end of grade two, this conclusion is less clear at the end of grade four. The per student annual costs for the Training and Coaching are roughly \$31 and \$43 respectively. Training thus increased the average oral reading proficiency (i.e. number of words read) by 25 per \$100 spent per student per year, compared to 21 the case of Coaching, at the end of grade four.⁹

Cost-effectiveness should also consider the benefits on future cohorts of students. When looking at both cohorts of students, the per student cost halves (assuming the same number of pupils in each cohort), but the average treatment effect is the average of the two separate treatment effects. With this comparison, Training remains more cost-effective with an increase of 37 words read per \$100 spent per student per year, compared to 31 in the case of Coaching.

⁹ To allow for comparisons across cohorts, we do cost-effectiveness analysis on the same sample of 164 schools.

Given the confidence intervals around estimated treatment effects and expected future costs, the more reasonable conclusion is that the two interventions are similarly cost-effective. Even though Training is slightly more cost-effective, as measured by the ratio of effect size to cost, this must be weighed against the greater risk of not having any positive impact in the second year, since we cannot reject the null at a 90% confidence that it has no impact on our aggregate score of reading proficiency.

Results on teacher behavior

Our results above demonstrate a positive impact of Coaching on the next cohort of students, one year after the teachers received Coaching. This implies at least some persistence in teacher knowledge, application of teaching techniques, and/or use of resources in the classroom. In this section, we more directly test for this sustained impact. Results from the teacher surveys, classroom observation, and document inspection reveal a consistent trend that the teachers in both treatment arms still have access to and use the resources available to them, and they demonstrate higher pedagogical knowledge. However, teachers that received Coaching use these resources more intensively. Moreover, for teaching practice, print-richness of the environment, and students' use of exercise books, the effects are only statistically significant for teachers that received Coaching. This suggests that persistent behavior change requires a more intensive intervention of ongoing monitoring and feedback, and cannot be achieved through training alone.

As a starting point, Figure 5 shows that only 68 percent of grade three teachers in the control were teaching the same grade at the same school the previous year, compared to 77 percent of teachers in the Training and Coaching arms, respectively. It is thus possible that we are under-estimating the true effect of these programs on learning, since it is not capturing the benefits from trained/coached teachers who have left the schools or are teaching in another grade. Our main estimates on learning should therefore be interpreted as an Intent-to-Treat.¹⁰ For all subsequent analyses we restrict the sample to the 162 teachers (from 126 schools) who were teaching the same grade at the same school the previous year. To maximize comparability, we also restrict the sample of grade two teachers surveyed in 2015 to the same set of schools.

Next, Figure 6 shows that, although teachers in the evaluation arm were more likely to have received professional in-service teacher training on teaching home-language literacy in 2017 (the year of intervention), teachers that received Coaching were *less* likely to receive training the subsequent year. It is possible that government re-allocated training resources away from the schools that received Coaching, or that school principals assigned other teachers in the school to receive professional development that

¹⁰ We do not instrument treatment on the probability of being treated and in the same grade and school the next year, because there could be positive spillovers of the program to the rest of the school.

year.¹¹ Once again, this would imply an underestimate of the treatment effects on learning, if these other professional development activities also contributed to learning. As a comparison, close to 100% of the grade two teachers stated that they had received training or support.

Table 5 shows that for both Training and Coaching there is a persistent effect on teachers' knowledge. Teachers in the Training and Coaching arms are 28 and 41 percentage points more likely to answer a question about teaching phonics correctly.¹² As a comparison, in column (2) we show the impact on our sample of grade two teachers, who were assessed the same year that they were exposed to the program. The magnitudes of the coefficients are roughly equivalent, suggesting no deterioration in knowledge. We cannot rule out, of course, that there was deterioration in knowledge of more sophisticated pedagogical questions.

Table 5, columns (3) to (5), and Figure 7 show high persistence in the use of resources. 55 and 35 percent of teachers in the Training and Coaching arms, respectively, used lesson plans provided by an NGO, compared to only 13 percent in the Control. Similarly, 98 and 94 percent of teachers in the two treatment arms make use of the graded readers, compared to only 12 percent in the control. For comparison, when we surveyed the grade two teachers, roughly 90 percent of teachers in the intervention arms were using the lesson plans, and 95 percent were using the graded readers. This suggests some deterioration in use of the lesson plans, yet a persistent statistically significant treatment effect. Moreover, column (5) and Figure 7 show that teachers that received that received Coaching are far more likely to make daily use the graded readers. These resources are supposed to take place during group-guided reading, which is supposed to take place daily. They therefore use the resources more *intensively*, and at the appropriate level.

Consistent with this result, Table 6 shows that the impacts on teacher *behavior* mostly fade out for teachers that received Training but persist for teachers who received Coaching. The grade two teachers – i.e. those who were assessed the same year as their participation in the programs – were 17.9 and 24.6 percentage points more likely to implement teaching activities at the correct weekly frequency. This effect is lower for the sample of grade three teachers, who were surveyed the year following the intervention, and is no longer statistically significant for teachers who received Training. This same pattern – i.e. persistence for teachers that received Coaching, but fade-out for teachers that received Training – is also present for our index of the print-richness of the environment, and student completion

¹¹ The research team visited the provincial government multiple times to inform them of the study and report results. The research team also communicated the most recent results of the evaluation to all the schools in the sample, during the third round of data collection. The program was therefore well-known by all the schools in our evaluation sample, as well as the provincial government.

¹² We asked teachers how many times a sound should be repeated. The correct answer is 3.

of exercises in their exercise book and government handbook. The print-richness of the environment is an indicator of teaching activity and effort. Even though teachers in both programs received print to place in their classroom, such as posters and flashcards, these resources deteriorate quickly, and the teacher needs to maintain them. Moreover, the flashcards displayed in the classroom need to be regularly updated depending on the content being taught. Similarly, increased usage of the exercise books means that the teacher provided students with exercises to practice – something which requires effort and preparation from the side of the teacher.

Tables A.2 to A.4 in the appendix unpack these three indices to see which indicators are driving the results. For teachers that received Coaching, there is persistence for three of the five activities - group-guided reading, shared reading, and writing – and for all the indicators of print-richness of the classrooms. With pupil use of exercise books, there is also a large effect for teachers that received Training on days completed any exercise in the exercise book. The impacts are statistically significant for the more challenging exercises that required writing paragraphs. However, students of teachers who received Training write less frequently in the government-issued handbook. There thus seems to be substitution away from using the government-handbook for teachers that received Training, but not for Coaching.

Note that since the measures of teaching activities are self-reported, it is possible that the results are driven by social desirability bias, if teachers merely give what they perceive to be correct answer to impress the fieldworker. But there are three reasons why this is unlikely to be driving the results. First, such an interpretation needs to explain why teachers' propensity to lie in order to impress the fieldworker fades out for teachers that received Training, but not Coaching. Second, Cilliers et al (2018) found consistent results on teaching practice between the survey measures and lesson observations. Third, one would expect teachers to state that they are more likely to implement *all* activities if they merely answer to impress the fieldworker, but this is not the case. For example, coached teachers actually report to teach phonics less frequently than the control, because it is not supposed to take place on a daily basis. Finally, even in the extreme case that *all* teachers lie, this remains a credible indicator of knowledge, since the teacher needs to know what the correct answer is.

Table 7 shows that there is persistence in knowledge and use of resources even for grade one teachers, who were exposed to the program three years ago. This sample is restricted to 100 grade one teachers who were teaching in the same school and same grade in 2015 – the year that they received the training/coaching. Most strikingly, almost all teachers in the two treatment arms still use the *Vula Bula* graded readers, compared to only 9 percent in the control. But there is no persistence in the use of exercise books and print-richness of the environment.

Table 8 tests for school-level spillovers: the impact on teachers who were not directly exposed to the program, but still teach in the treated schools. The sample is restricted to teachers who were not teaching in the same grade and in the same school the year that intervention took place (i.e. grade three teachers who were not teaching grade three at the same school in 2017, and grade one teachers who were not teaching grade one at the same school in 2015). The regression also includes a dummy for the grade of the teacher. Even teachers who were not directly exposed to the program are more likely to use the graded readers, use the lesson plans, and demonstrated improved pedagogical knowledge. Though there are no positive significant impacts on teaching activities, the print-richness of the environment, or use of exercise book. Clearly, schools find the resources useful and new teachers continue using them, even if this does not necessarily lead to improve teaching practices.

To summarize, these results show persistent impacts in teacher knowledge and access to resources, up to three years after teachers were exposed to the program. However, there is evidence of mild deterioration in teaching *practice*, and this deterioration is largest in the Training arm, relative to Coaching.

Conclusion

This study tests for the sustainability of two teacher professional development programs along two dimensions: persistence in learning gains for the original cohort of students, and persistence in teaching productivity. Encouragingly, there is no evidence of fade-out of results for the first cohort. However, there is a substantial reduction in teacher productivity. The impacts on learning for the second cohort of students – i.e. who were taught by teachers one year after the program terminated – is far lower, and only statistically significant in the Coaching arm.

The sustained impact on subsequent cohorts dramatically changes decisions around cost-effectiveness and scalability of coaching interventions (Kraft et al, 2018). Although it is prohibitively costly for high-quality coaches to visit all schools at the same time, one approach that policymakers could consider is to rotate teams of coaches to different schools. The above results suggest that changes in teaching practice can be sustained up to three years later, but that impacts on learning do diminish somewhat with time. This would therefore provide support for a rotation model of scaling up coaching programs.

At the same time, the fact that the impacts roughly halved the second year suggest the need for additional ongoing investment in the teachers to sustain the improved teaching practices. Given the fact that the fixed costs are already sunk, a model of periodic visits by coaches in subsequent years could be highly cost-effective. It is also possible that more prolonged exposure to training or coaching would lead to stronger and more persistent impacts, if teachers are given more opportunities to develop their skills.

Bibliography

Baird, Sarah, et al. "Worms at work: Long-run impacts of a child health investment." *The Quarterly Journal of Economics* 131.4 (2016): 1637-1680.

Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2019). How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, 0618-9538R

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. "Labor market returns to an early childhood stimulation intervention in Jamaica." *Science* 344, no. 6187 (2014): 998-1001

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. "Broken or fixed effects?." *Journal of Econometric Methods* 8.1 (2018)

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic inquiry*, 46(3), 289-324

Heckman, James J., et al. "The rate of return to the HighScope Perry Preschool Program." *Journal of public Economics* 94.1-2 (2010): 114-128

Lin, Winston. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7.1 (2013): 295-318.

Majerowicz, Stephanie, and Ricardo Montero. "Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru." (2018)

Tables and Figures

Table 4. Impacts on learning for the second cohort

VARIABLES	(1) Aggregate	(2) Letters	(3) Words	(4) Paragraph	(5) Comprehension	(6) Written comprehension
Training	0.113 (0.0975)	-0.724 (2.166)	0.975 (1.505)	3.754 (2.529)	0.0361** (0.0174)	0.0278 (0.0228)
Coaching	0.167* (0.0982)	0.956 (1.774)	1.461 (1.533)	4.330 (2.699)	0.0531*** (0.0196)	0.0383 (0.0237)
Observations	1,618	1,636	1,636	1,636	1,636	1,632
R-squared	0.130	0.101	0.117	0.127	0.083	0.099
Control mean	0	42.71	25.90	38.13	0.235	0.246

Notes: Each column represents a separate regression, using equation 1. All specifications include the following controls: gender, age, school performance in standardized national exam, a district dummy, a community-level wealth index and highschool attendance rates. Standard errors are in parentheses and clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Figure 1: The “Matthew effect” hypothesis. The pace of learning is positively correlated with the current stock of knowledge.

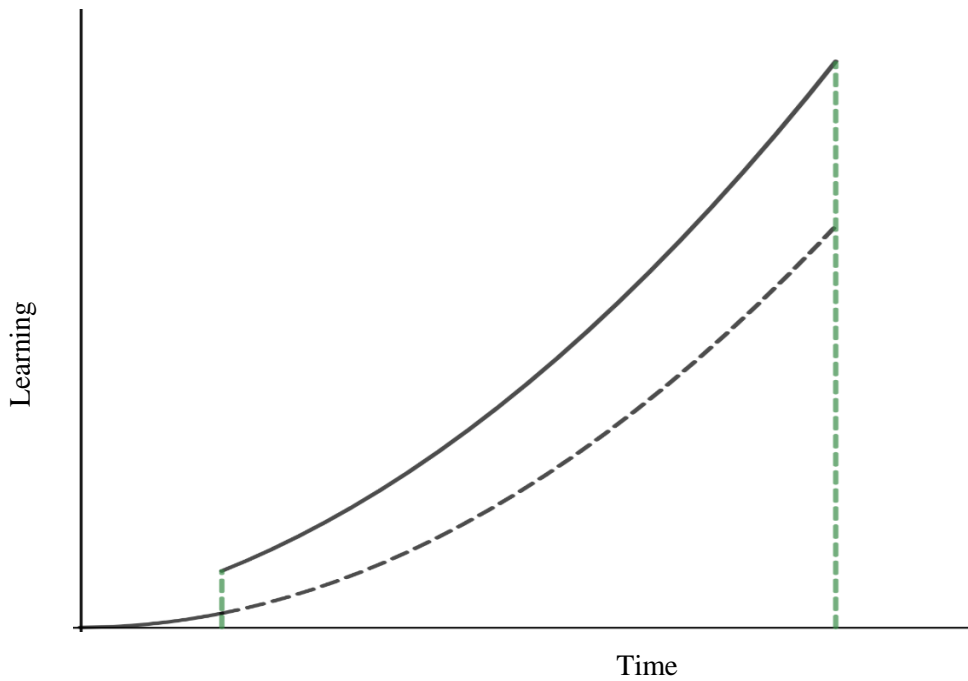


Figure 2: The “development lag” hypothesis. An early intervention enables students to learn faster at the start, but students in the control eventually catch up.

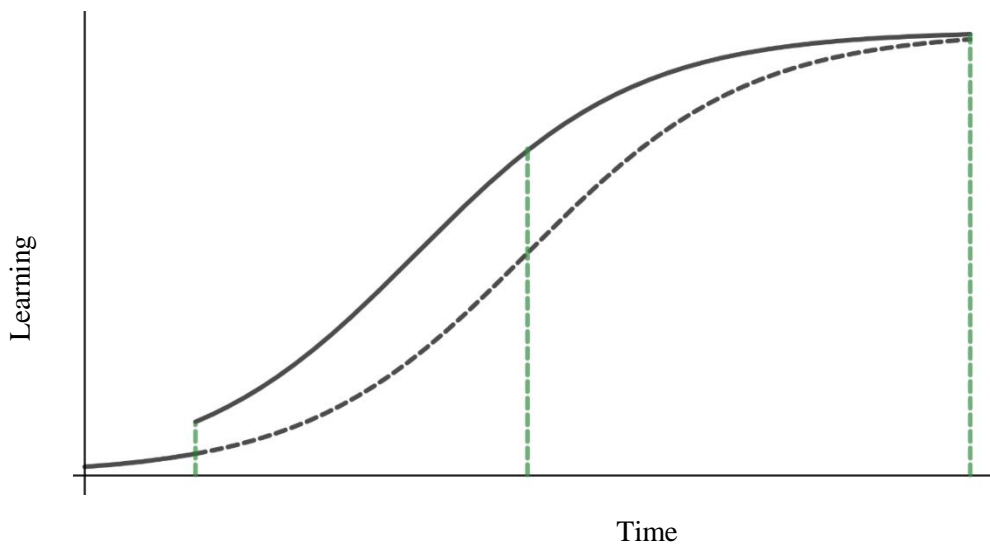


Figure 3: Attrition and grade progression by evaluation arm

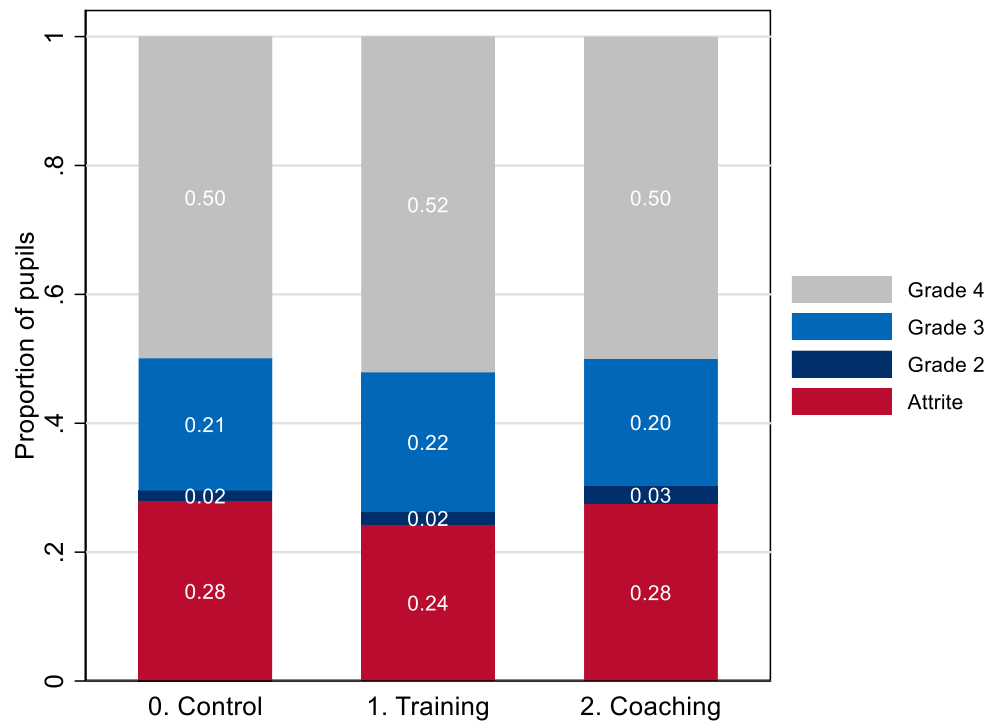


Figure 4: Paragraph reading by grade and treatment arm

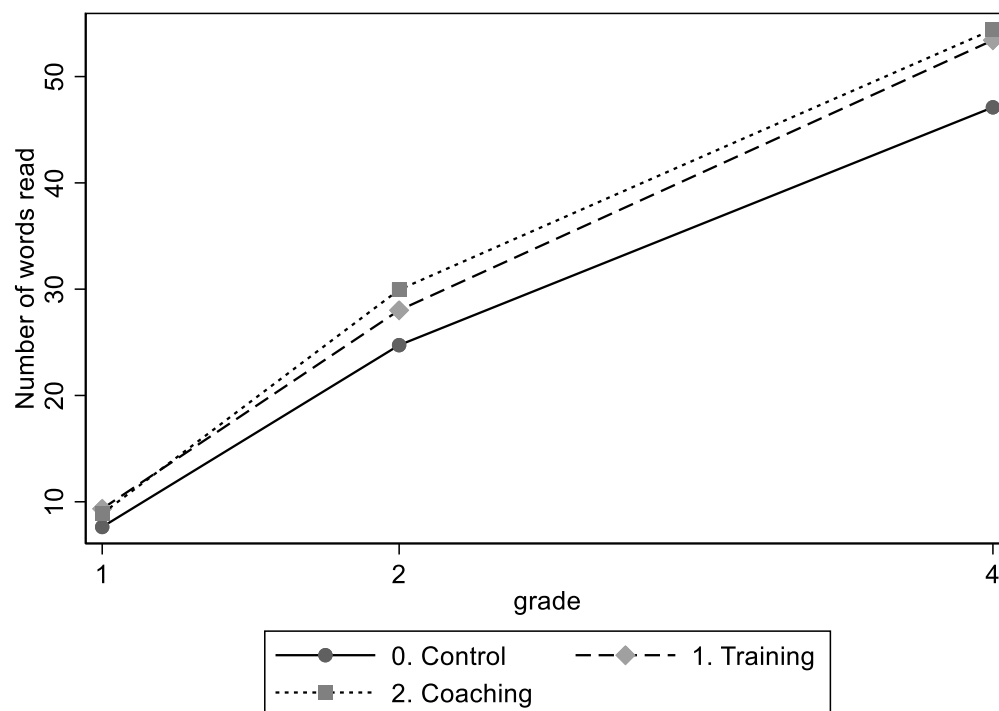
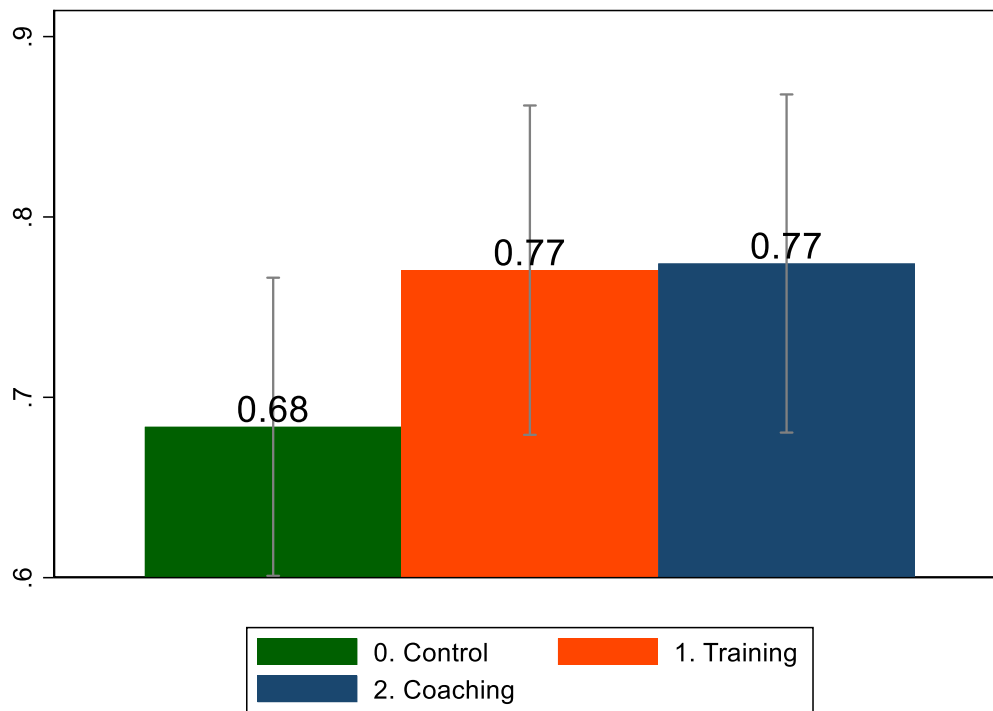


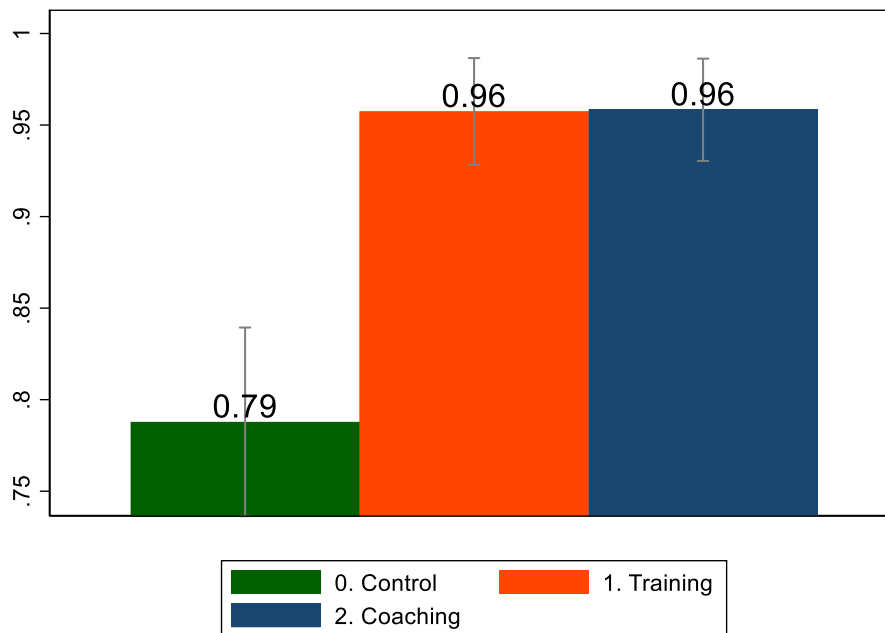
Figure 5: The probability of teaching the same grade and same school the subsequent year



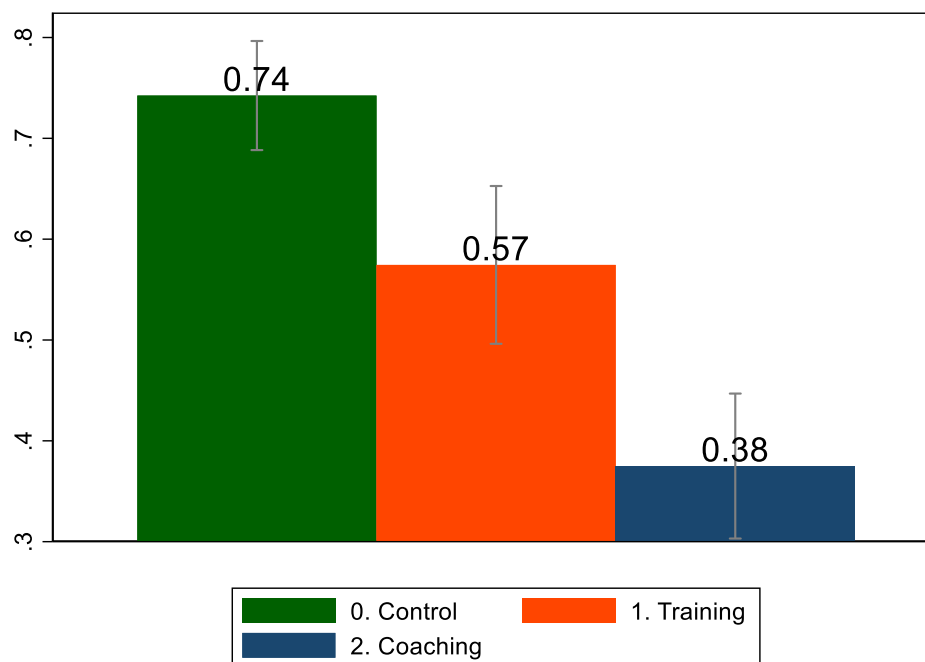
Note: The lines show a 90 percent confidence interval, with standard errors clustered at the school-level. Sample restricted to grade 3 teachers.

Figure 6: Proportion of grade three teachers that received training

Panel A. In 2017 (year of intervention)



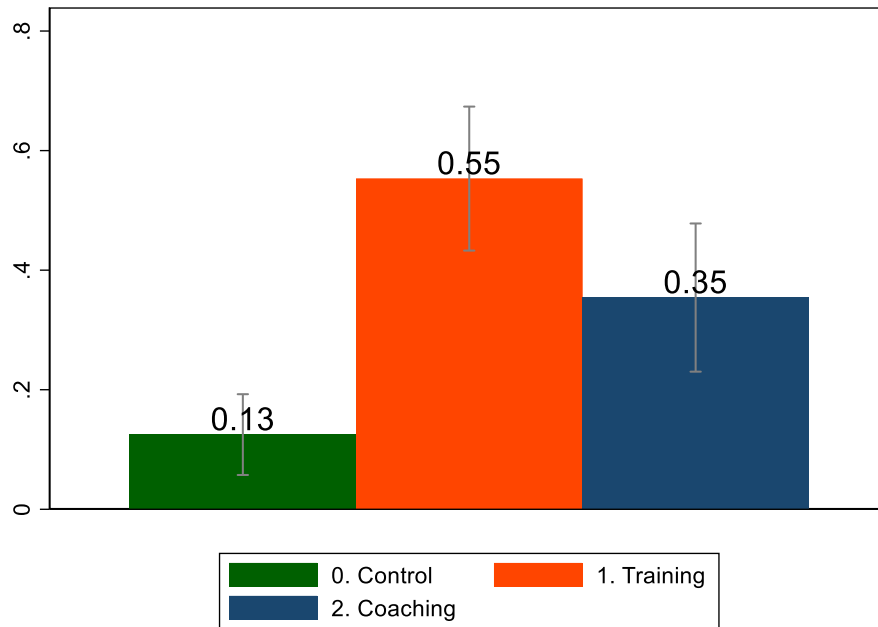
Panel B. In 2018 (year after intervention)



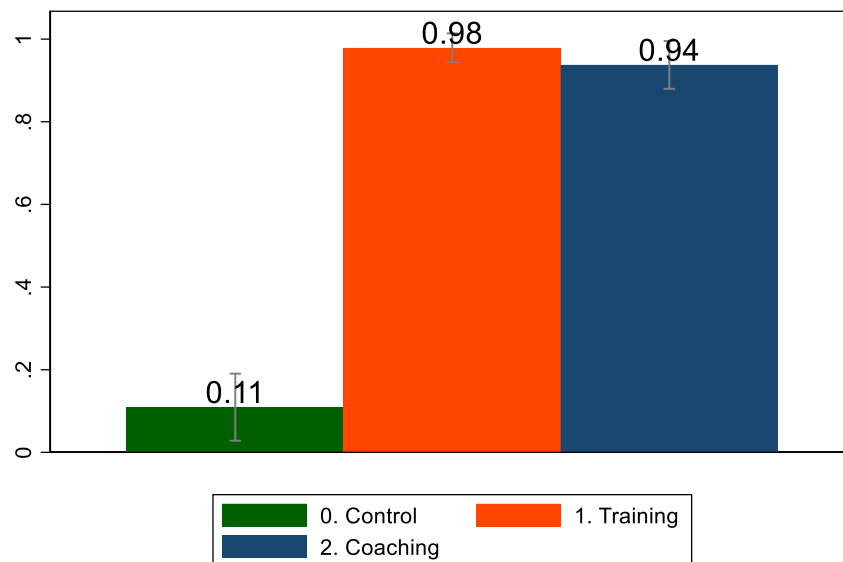
Note: Data restricted to grade three teacher that were teaching grade three at the same school the previous year. The lines show a 90 percent confidence interval, with standard errors clustered at the school-level.

Figure 7: Use of resources

Panel A. Use lesson plans provided by an NGO



Panel B. Use the *Vula Bula* graded readers.



Note: Data restricted to grade three teacher that were teaching grade three at the same school the previous year. The lines show a 90 percent confidence interval, with standard errors clustered at the school-level.

Table A.1.Comparing the impacts on the first and second cohorts

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	Aggregate			Paragraph			Comprehension			Written comprehension
Training	0.213*** (0.0768)	0.216*** (0.0782)	0.113 (0.0975)	5.143*** (1.783)	7.854*** (2.372)	3.754 (2.529)	0.0494* (0.0250)	0.0474*** (0.0174)	0.0361** (0.0174)	0.0535 (0.112)
Coaching	0.347*** (0.0740)	0.286*** (0.0691)	0.167* (0.0982)	7.956*** (1.563)	8.823*** (2.110)	4.330 (2.699)	0.104*** (0.0243)	0.0595*** (0.0161)	0.0531*** (0.0196)	0.346*** (0.1000)
Grade	2	4	3	2	4	3	2	4	3	4
Cohort	1st	1st	2nd	1st	1st	2nd	1st	1st	2nd	1st
Observations	2,728	2,200	1,618	2,728	2,200	1,636	2,728	2,200	1,636	2,200
R-squared	0.209	0.194	0.130	0.184	0.187	0.127	0.153	0.153	0.083	0.070
Control mean	0	0.0146	0	24.48	47.79	38.13	0.309	0.304	0.235	1.535

All results are estimated on the same sample of 164 schools.

Table A.2. Teaching activities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Group-guided reading		Phonics		Shared reading		Spelling		Writing	
Training	0.167** (0.0836)	0.0928 (0.0979)	-0.0516 (0.0837)	-0.0845 (0.0930)	0.247*** (0.0783)	0.0709 (0.0954)	0.156** (0.0664)	0.0758 (0.0664)	0.353*** (0.0845)	0.0656 (0.0990)
Coaching	0.201** (0.0804)	0.184* (0.0992)	0.192** (0.0830)	0.0800 (0.0960)	0.235*** (0.0815)	0.250*** (0.0800)	0.209*** (0.0556)	0.0754 (0.0603)	0.391*** (0.0832)	0.220** (0.100)
Years after treatment	0	1	0	1	0	1	0	1	0	1
Teacher grade	2	3	2	3	2	3	2	3	2	3
Observations	207	162	207	162	207	162	206	162	207	162
R-squared	0.078	0.105	0.091	0.072	0.111	0.125	0.108	0.074	0.153	0.107
Mean in control	0.235	0.313	0.469	0.522	0.383	0.493	0.725	0.836	0.272	0.433
P-value Training = Coaches	0.713	0.411	0.00688	0.114	0.885	0.0620	0.389	0.995	0.667	0.158

Note. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Note: For grade 3 teachers the sample restricted to teachers that also taught grade three in the same school the year before.

Table A.3. Print richness of the classroom

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Reading corner		Story books		Setswana wall charts		Picture wall charts		Flash cards	
Training	0.263 (0.186)	-0.103 (0.196)	0.204 (0.184)	-0.0898 (0.193)	0.439** (0.213)	-0.0261 (0.193)	0.465* (0.257)	0.0875 (0.200)	0.738*** (0.217)	0.320 (0.201)
Coaching	0.410** (0.204)	0.426** (0.192)	0.203 (0.154)	0.162 (0.176)	0.485*** (0.184)	0.340** (0.166)	0.252 (0.209)	0.435** (0.196)	0.751*** (0.206)	0.689*** (0.195)
Years after treatment	0	1	0	1	0	1	0	1	0	1
Teacher grade	2	3	2	3	2	3	2	3	2	3
Observations	190	160	197	160	197	160	198	160	197	160
R-squared	0.094	0.112	0.064	0.095	0.078	0.144	0.101	0.152	0.201	0.217
Mean in control	2.592	2.470	3.154	3.258	2.747	2.545	2.582	2.364	2.410	2.364
P-value Training =										
Coaches	0.466	0.00796	0.995	0.263	0.822	0.0684	0.380	0.118	0.947	0.0651

Note. See table 6.

Table A.4., Number of days completed a workbook, and number of pages completed in the exercise book.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Number of days completed - Learner exercise book							Government Handbook	
VARIABLES	Any		Writing		Full sentence		Paragraph	Prop.	No.
Training	19.18*** (3.979)	9.189 (6.517)	9.275*** (3.214)	7.728 (6.253)	12.58*** (3.872)	6.241 (4.848)	2.353 (1.438)	0.0343 (0.0559)	-4.565** (2.247)
Coaching	10.07** (4.570)	16.35** (7.855)	9.829** (4.135)	13.02* (7.323)	7.467** (3.762)	6.908* (3.792)	3.169* (1.831)	0.118*** (0.0427)	-0.0300 (2.212)
Yrs after treat	0	1	0	1	0	1	1	0	1
Teacher grade	2	3	2	3	2	3	3	2	3
Observations	203	156	203	156	203	156	156	193	154
R-squared	0.154	0.094	0.139	0.092	0.129	0.064	0.086	0.064	0.072
Mean in control	21.99	34.52	17.32	32.67	13.52	16.92	5.030	0.789	18.98
P-value Training = Coaches	0.0810	0.401	0.906	0.514	0.252	0.903	0.659	0.0723	0.0433