Monitoring Teachers and Changing Teaching Practice: Evidence from a Field Experiment.[*]

Jacobus Cilliers,[†] Stephen Taylor[‡]

August 2017

Abstract

There is large documented variation in teacher quality, which strongly correlates with teaching activity in the classroom. But how to change teaching practice? We present results of a randomized evaluation of two different programs aimed at improving early-grade reading in poor schools in South Africa. Both programs provide teachers with structured lesson plans and supporting reading materials, but they differ in the mode of implementation. In some schools, teachers receive two two-day training sessions over the course of the year. In other schools, teachers receive monthly visits from specialized reading coaches who monitor teaching activity and provide feedback. The curriculum and expected teaching methods are the same as in the control. We track a cohort of pupils over two years and find that over this period the coaching program improved learning proficiency by 0.25 standard deviations. In contrast, the training program had a far smaller and statistically insignificant impact of 0.11 standard deviations. The impacts are also dramatically larger in urban schools: 0.42 and 0.75 standard deviations in the training and coaching arms respectively. We conclude that, under a variety of different assumptions, coaching is more cost-effective. Moreover, data from detailed lesson observations reveal that teachers in both interventions are more likely, relative to the control, to enact group-guided reading— a difficult teaching technique where children read aloud in small groups sorted by ability. Related to this technique, intervention teachers are able to provide more individual attention and pupils are more likely to read from reading booklets. These changed classroom practices are consistently larger for teachers that received coaching. Results of this paper yield insights into the appropriate combination of support required to shift teaching practice, and which practices plausibly have a greater impact on learning.

[*] We are grateful for useful comments and feedback from Servaas van der Berg, and fantastic research and management support from Janeli Kotze and Mpumi Mohohlwane. The International Initiative for Impact Evaluation provided funding for the evaluation. All errors and omissions are our own.

[†] McCourt School of Public Policy, Georgetown University

[‡] Department of Basic Education, South Africa

1 Introduction

How to change teaching practice? A large body of evidence suggests that teachers play a critical role in shaping children's learning (Rivkin, Hanushek & Kain 2005, Kane & Staiger 2008, Araujo, Carneiro, Cruz-Aguayo & Schady 2014). The difference between a bad and a good teacher matters greatly, not only for how much a child learns in a year, but also for his/her future academic performance and employment prospects (Chetty, Friedman & Rockoff 2014).

Moreover, although observed teacher characteristics —such as experience and education— often do not predict teacher quality (Rivkin et al. 2005), classroom practices do. Results from classroom observations show large variation in teaching practice, even within the same school, and these practices correlate with faster learning trajectories (Allen, Gregory, Mikami, Lun, Hamre & Pianta 2013, Kane & Staiger 2012). In short, teachers matter a great deal, but it depends less on who the teacher is than what s/he does in the classroom.

Taken together, these results (optimistically) suggest large potential learning gains by improving teaching practice, possibly through developing the capacity of the existing pool of teachers. This is encouraging, because efforts to motivate teachers to increase their effort levels — either through financial incentives (Glewwe, Ilias & Kremer 2010, Muralidharan & Sundararaman 2009, Cilliers, Kasirye, Leaver, Serneels & Zeitlin 2014) or increased accountability (Duflo, Dupas & Kremera 2011, Muralidharan & Sundararaman 2013)— are often not politically tractable.[1] Furthermore, these same political constraints mean that it is near impossible in most countries to get rid of bad public sector teachers.

Indeed, lots of resources are already spent by government and donors on improving the quality of the existing teachers, most commonly through in-service teacher training. For example, by some estimates the United States spends $18,000$ USD per teacher on professional development (Jacob & McGovern 2015, Fryer 2017).[2] Over 78 per cent of South African grade 2 teachers received in-service teacher training in 2016.[3] And Evans and Popova (2016) estimate that nearly two thirds of World Bank-funded education programs include a professional development component.

But the evidence on the impact of in-service teacher training is mixed at best. Many studies in the United States have found no impact of professional development programs on student learning (Harris & Sass 2011, Garet, Wayne, Stancavage, Taylor, Eaton, Walters, Song, Brown, Hurlburt, Zhu et al. 2011, Garet, Cronen, Eaton, Kurki, Ludwig, Jones, Uekawa, Falk, Bloom, Doolittle et al. 2008, Jacob & Lefgren 2004). In a recent meta-analysis of evaluations of in-service teacher training programs in developing countries, the authors concluded that "rigorous evidence on the effectiveness of such programs remains" limited (Popova, Evans & Arancibia 2016).

A plausible reason for the failure of many training programs is that they commonly focus on changing knowledge (often through one-off workshops at a central venue), yet teaching is a skill that needs to be

[1] For example, recent education reform in Indonesia attempted to put in place pay scale differentials based on performance metrics, but it was blocked by trade unions. The eventual policy lead to an unconditional doubling of teacher salaries, with no resultant impact on learning (De Ree, Muralidharan, Pradhan & Rogers 2015).

[2] Professional development can be broadly defined as any activity that helps existing teachers improve their teaching knowledge and capacity. It often entails workshops, but could also include regular follow-up from mentors and online self-learning courses. For the purpose of this study, I refer to in-service teacher training and teacher professional development interchangeably.

[3] Based on the teachers in the control group of our evaluation sample

developed through ongoing practice. New teaching techniques are often difficult to enact at first.[4] But with regular practice, the teacher can become more adept in a difficult technique and, as a result, find it easier to implement.

A potentially cost-effective way to encourage enactment of difficult teaching techniques is to combine training with structured lesson plans. Lesson plans reduce the cost of transition since they require no additional lesson preparation. The structure also provides a regular routine, which facilitates practice and fosters habits, so that the activities eventually require minimal conscious effort to enact (Neal, Wood & Quinn 2006). In the United States, for example, 'off-the-shelf' lesson plans for mathematics teachers improved students' math achievement (Jackson & Makarin 2016).

But training with lesson plans might not be sufficient to change behavior without ongoing observation and feedback from someone who has already mastered the skills —a coach. Some techniques might remain too difficult or arduous to implement. Worse, the new technique might be incorrectly applied and bad habits could form.[5] An expert coach, who observes the teacher in the classroom, can provide feedback to rectify mistakes and improve on techniques. The coach can also demonstrate correct teaching techniques, and provide additional motivation for the teacher to change behavior.

Consistent with this, four studies from three different countries —Liberia, Kenya, and Uganda— have demonstrated that a bundled intervention of training, lesson plans, and coaching can dramatically improve pupils' proficiency in early-grade reading (Piper, Zuilkowski & Mugenda 2014, Piper & Korda 2011, Lucas, McEwan, Ngware & Oketch 2014, Kerwin, Thornton et al. 2015). But there is much we still do not know. We do not know if some components are uniquely responsible for the learning gains, and if a 'light-tough' version could be more cost-effective. In addition, the programs typically introduce a new curriculum, so we do not know whether to attribute the impacts to new teaching content or changed teaching practice. And without classroom observations, one cannot tell if and how teaching practice changed.

Is centralized training combined with structured lesson plans sufficient to change teaching practice and ultimately improve learning, or should it be combined with ongoing monitoring and feedback from coaches to be effective? Which method is more cost-effective, and for which types of learning outcomes? How does the impact depend on the teaching environment and/or baseline teacher quality? Which teaching practices are hardest to shift, and which practices matter most to improve learning?

To answer these questions, we conduct a randomized evaluation of two different interventions aimed at improving home-language early-grade reading in poor communities in South Africa,[6] randomly assigning 50 schools to each intervention and 80 schools to the control.[7] Both interventions provide teachers with structured lesson plans and supporting materials such as graded reading booklets, flash cards, and posters. The lesson plans are integrated with government curriculum and mirror exactly the pedagogical techniques prescribed by government.[8] Both treatment and control schools also follow the same

[4] The challenge is amplified by the fact that in-service teacher training necessarily requires teachers to *change* their behavior, rather than learn new behavior. This requires abandonment of an old approach that teachers were comfortable in doing (Kennedy 2016).

[5] Moreover, lesson plans reduce teacher autonomy and could thus hinder a good teacher's ability to cater his/her teaching to the needs of the child.This is particularly concerning if the lesson plans follow an overly-ambitious national school curriculum —as is arguably the case in most developing countries.

[6] We sampled schools in the bottom three quintiles, in terms of socio-economic status

[7] A third intervention focused on parental involvement, but is not the focus of this paper and will be part of an upcoming paper.

[8] The pedagogical standards go as far as to prescribe weekly intensity of enacting different teaching techniques. For

government handbook that specifies content that needs to be covered each day.

The two interventions differ in the level of support teachers receive to implement the lesson plans. The first intervention (which we refer to as Training), trains teachers on how to use the lesson plans through central training sessions, each lasting two days, and occurring twice yearly. This mode of implementation is similar to most government in-service programs. In the second intervention (which we refer to as Coaching), specialist reading coaches visit the teachers on a monthly basis to provide ongoing feedback and support. Coaching costs roughly 62 USD per pupil, compared to 48 USD for Training. We assessed the reading ability of a random sample of 20 pupils in each school at three points in time: once as they entered grade one prior to the roll-out of the interventions (February 2015), and again at the end of their first and second academic years (November 2016 and 2017 respectively). During these school, visits we also surveyed teachers and the school principal. Since pupils are typically taught by a different teacher in their second year, we mostly surveyed a different set of teachers in the second year. Finally, we conducted detailed lesson observations and in-depth teacher surveys in a (stratified) random sample of 60 schools in October 2016— 20 schools in each evaluation arm.

We find that, after two years of exposure to the program, pupils' reading proficiency increased by 0.25 standard deviations relative to the control if their teachers received ongoing mentoring from reading coaches. In contrast, the impact of Training is far smaller (0.11 standard deviations) and is not discernible from zero given conventional levels of significance. The impacts are larger when we exclude multi-grade classrooms, a setting where the program was never intended to work. We also find dramatic geographic variation in effective sizes. Both programs were substantially more effective in urban schools, with an increase of 0.75 and 0.42 standard deviations for the Coaching and Training programs respectively. Given these estimates, we can conclude that Coaching is more cost-effective than Training with an estimated 0.41 standard deviation increase in reading proficiency per 100 USD spent, compared to 0.23 in the case of Training. This conclusion remains unchanged if we assume that the program will only be scaled up in urban schools.

Turning to teaching practice, two results are worth emphasizing. First, teachers in both treatment arms are more likely to practice group-guided reading— an activity where the pupils read aloud in smaller groups that are ideally sorted by ability. Government curriculum requires that this takes place on a daily basis. Group-guided reading forces pupils to engage individually with the text, rather than mimic what the teacher is reading with the class as a whole. It also opens up the possibility for individual feedback from a teacher, as s/he can now move between groups. Furthermore, if different groups are assigned different graded reading booklets based on ability, this allows pupils to progress at a pace appropriate to their level of proficiency. Consistent with this, the lesson observations reveal that pupils in both the treatment arms are more likely to read aloud in groups and receive individual attention from the teacher when they read. The groups are rarely sorted by ability, though.

The impacts on group-guided reading are consistently stronger for teachers who received coaching. Notably, we see no change in other activities that are also required to take place daily, but are easier to implement: teaching of phonemic awareness and letter recognition.[9] This suggests that this technique

example, group-guided reading and phonics should take place on a daily basis. Spelling tests should take place on a weekly basis.

[9] Phonemic awareness and letter recognition is taught through whole-class reading, where all the children in the classroom follow or read with the teacher.

requires additional coaching for some teachers who find it to difficult to implement.[10]

Second, even though there is no large difference between treatment arms and the control in access to the graded reading booklets, the lesson observations reveal that far more pupils are reading them in the program schools. Again, this difference is much larger in the Coaching arm, and larger still for coached teachers from urban schools. Strikingly, virtually no pupils in the control are reading the reading booklets. The graded reading booklets are meant to be read during group-guided reading. These results thus reveal the important interaction between resources, teaching practice, and use of resources: provision of reading material is insufficient if teachers do not apply the appropriate teaching techniques so pupils have opportunities to use them.

Taken together, our results show that the 'light-touch' combination of training and lesson plans can shift teaching practice, but the shift is far larger when teachers receive ongoing monitoring and feedback from a coach. Moreover, our results provide strong suggestive evidence for which classroom activities matter most for learning: providing pupils with more opportunities to practice reading, and more individual feedback from the teacher. This seems like a trivial result, but in large classrooms providing individual attention is a non-trivial task.

This conclusion is in line with the growing body of evidence from developing countries that pedagogy that targets teaching to the level of the child can be highly effective at improving learning (Evans & Popova 2015). Randomized evaluations have found that remedial education programs (Banerjee, Cole, Duflo & Linden 2007), additional teacher assistants (Duflo & Kiessel 2014), or contract teachers (Duflo et al. 2011) can improve test scores since they free up resources to provide additional attention to worse-performing pupils. What is encouraging from our study is that we show that individual attention can be accomplished by changing pedagogical practices of the existing pool of teachers, without additional teachers or computer technology.

The paper proceeds as follows: section 2 describes the interventions and the motivating theoretical channels, section 3 describes the evaluation design and empirical strategy, section 4 reports the results, and section 5 concludes.

2 Program description

Working with the non-governmental organization, Class Act, we designed two related interventions aimed at improving early-grade reading in one's home language. The home language of interest, Setswana, is spoken by the majority of the population in the province where the program was implemented, North-West province. Both interventions provide teachers with structured lesson plans, which provide detailed specification for each lesson, including information on methodology and content to be taught for each instructional day. In addition, teachers receive supporting materials, such as graded reading booklets, flash cards, and posters. The graded reading booklets provide a key resource for the teacher to use in group-guided reading (discussed in more detail below) and individual work so as to facilitate reading practice at an appropriate pace and sequence of progression.

The two interventions differ in the level of support and training teachers receive to implement the

[10] In the in-depth teacher surveys, some treated teachers complained that group-guided reading was too difficult to implement in the classroom.

lesson plans. The one intervention trains the teachers on how to use the lesson plans and accompanying materials through central training sessions, each lasting 2 days and occurring twice yearly. We refer to this intervention as Training. The second intervention, which we refer to as Coaching, provides exactly the same set of instructional materials. However, instead of central training sessions, specialist reading coaches visit the teachers on a monthly basis to provide ongoing feedback and support. In addition to these on-site visits, there are occasional meetings with the coach and a small cluster of nearby schools that are part of this intervention.

It is important to note that the treated teachers are coached and trained in the same curriculum as teachers in the control. The lesson plans follow exactly the official government curriculum and its accompanying standards for teaching activity. These standards are very detailed and go as far is specifying the weekly frequency with which different teaching activities should take place. The lesson plans are also integrated with the government-provided workbooks, which detail daily exercises to be completed by students. Moreover, the control teachers also receive high level of support from government. For example, panels (a) and (b) in Figure 1 and rows (1) and (2) in Table 2 show that over 79 per cent of teachers in the control received in-service training on teaching Setswana as a home language the past year; and 96 per cent of teachers have at least some graded reading booklets in the classroom.

Any difference we observe is therefore due to the modality of support the teachers receive, not the pedagogical content.

2.1 Mechanism

In this sub-section we outline the pedagogical theory of reading acquisition that motivates the design of the curriculum and interventions, and we discuss the potential mechanisms through which lesson plans and coaches could change teaching behavior.

2.1.1 Group-guided reading using graded readers

An important building block to reading comprehension is decoding: the translation from symbols into speech. It is generally understood that this process develops in stages: phonemic awareness (awareness that speech is made out of separate sounds), letter recognition, word recognition, and eventually the ability to read sentences. Importantly, to reach a stage where one can automatically recognize print requires systematic and regular practice, with targeted feedback when pupils get it wrong.

One aspect of the curriculum, which is also reflected in the lesson plans, stands out as a way to facilitate practice and provide targeted feedback: Group-guided reading is "an ability-group reading teaching strategy where all the members in the group read the same text under the direction of the teacher." (Department of Basic Education, 2011). This is expected to take place on a daily basis and also requires regular assessment of pupils' reading proficiency in order to assign pupils to the appropriate graded reading booklets.

We posit three reasons why group-guided reading can be an effective method to teach reading. First pupils in smaller groups are forced to attempt decoding, because they cannot mimic the class as a whole.[11]

[11] In contrast, the most common (and easiest) type of teaching activity is whole-class reading where pupils jointly chorus sounds or read words with the teacher. But there is a risk that pupils merely mimic the sounds and words made by the teacher and their peers and are never forced to practice decoding themselves.

Second, pupils are more likely to receive individual feedback from a teacher, as the teacher can now move between groups. The teacher is thus able to identify and rectify incorrect reading. Third, if pupils are grouped by ability and different groups are assigned different graded reading booklets, then they can progress at a pace appropriate to their level of proficiency.

Note the relationship between resources and teaching activity: the reading booklets are necessary to effectively enact group-guided reading, but might have no impact on learning if they are not combined with group-guided reading.

2.1.2 Changing teaching practice

What is it about these program that could facilitate a change in teaching behavior? The theoretical basis that we find most convincing is that some important teaching techniques are difficult to enact, but become easier with regular practice. Whereas most traditional training programs impart knowledge, this program facilitates practice and fosters good teaching habits.

As a starting point, structured lesson plans reduce the cost of transition to a new pedagogical technique, since they require no additional lesson preparation. However, this might not be sufficient to shift behavior, if teachers find the new technique difficult. Indeed, many teachers in this study complained that they found it difficult to follow the lesson plans, especially the group-guided reading.

An additional benefit is that the structure of the lesson plans provide a regular routine which facilitates practice. The same teaching activities are repeated on a weekly basis and some activities, such as group-guided reading, are repeated on a daily basis. Moreover, the lesson plans provide useful prompts and reminders of what should be done in the classroom— activities that even the best-intentioned teacher might forget (Karlan, McConnel, Mullainathan, Zinman, ????).

With regular practice, teachers become more skilled at difficult techniques and so find them easier to do. At the same time, a regular routine breeds habits so that the activities requires minimal conscious effort to enact (Neal, Wood and Quinn, 2006). Both of these —the honing of a skill and formation of a habit— reduce the effort cost of enacting a difficult technique.

Turning to the reading coaches, they will clearly augment the practice required to master a new skill. Teachers are plausibly more likely to attempt the appropriate techniques when they are being monitored. The coaches can also identify incorrect applications, and provide feedback on how to improve. This reduces the risk of consistently applied new techniques incorrectly, and farming bad habits as a result. Furthermore, coaches will often also demonstrate a certain technique, so teachers also get an opportunity to observe an expert performing the activities.

3 Evaluation Design

3.1 Sampling and Random Assignment

The study is set in two districts in the North West Province, in which the main home language is Setswana. We chose this province, since it is relatively homogeneous linguistically and is one of the poorer provinces in South Africa, with many schools falling within the bottom three quintiles in terms of socio-economic status. Through a process of elimination, we developed a sampling frame of 230 eligible schools. In particular, we excluded schools that fall within the top two quintiles in terms of socio-economic status, schools where the language of instruction was not Setswana, schools that report to practice multi-grade teaching, schools that are extremely small or extremely large, and schools where we piloted the interventions. After all of these exclusions, 235 eligible schools remained. Using a random number generator, we then excluded 5 schools, which we retained as possible replacement schools.

Within this sample of 230 schools we created 10 strata of 23 similar schools based on school size, socio-economic status, and previous performance in the the national standardized exam, called the Annual National Assessments (ANA). Within each stratum we then randomly assigned 5 schools to each treatment group and 8 to the control group. All treatment schools with exception of one in the Coaching arm agreed to participate in the program. We included this school in the sample of treatment schools. Although the full evaluation consisted of three interventions (50 schools each), this paper only reports results for the two related interventions that provide scripted lesson plans. The third intervention —promotion of parental involvement— we will discuss in a separate paper.

3.2 Data collection

We visited each school three times: prior to the start of the interventions (February 2015), again after the first year of implementation (November 2015), and finally at the end of the second year (November 2016). During these school visits we administered four different survey instruments: A pupil test on reading proficiency and aptitude conducted on a random sample of 20 pupils who entered grade one at the start of the study, a school principal questionnaire, a teacher questionnaire, and a parent/guardian questionnaire. We assessed the same pupils in every round of data collection, but surveyed a different set of teachers between midline and endline, because pupils generally have different teachers in different grades. Finally, we also conducted lesson observations on a (stratified) random sub-set of 60 teachers in September 2016. The data-collection and data-capture organizations are independent and were blind to the treatment assignment.

We registered a pre-analysis plan at XX in October 2016, before we had access to the endline data.

3.2.1 Pupil assessment

The pupil test was designed in the spirit of the Early Grade Reading Assessment (EGRA) and was administered orally by a fieldworker to one child at a time. The letter recognition fluency, word recognition fluency and sentence reading components of the test were based on the Setswana EGRA instrument, which had already been developed and validated in South Africa. To this, we also added a phonetic awareness component in every round of assessment. The baseline instrument did not include all the same sub-tasks as the midline/endline instruments, because of different levels of reading proficiency expected over a two-year period. For baseline, we also included a picture comprehension (or expressive vocabulary) test since this was expected to be an easier pre-literacy skill testing vocabulary, and thus useful for avoiding a floor effect at the start of grade 1 when many children are not expected to read at all. Similarly, we included a digit span memory test —this involved repeating by memory first two numbers, then three, and so forth up to six numbers, and the same 5 items for sequences of words. The logic of including this test of working memory is that it is known to be a strong predictor of learning to read and would thus serve as a good baseline control to improve statistical power. For the midline and endline, we added a writing and a paragraph reading sub-task. We piloted and refined baseline instruments over 2014 and piloted and revised the midline/endline instruments over September 2015.

Out of the 3,539 pupils surveyed in baseline, we were able to re-survey 2,951 in midline, yielding an attrition rate of 16.6 per cent. The attriters had either moved school (1.6 per cent) or were absent on the day of assessment (14.4 per cent). Moreover, an additional 13% of our original sample are repeating grade one. Figure A.1 in the appendix shows the breakdown of groups. Column (1) in table A.1. regresses treatment assignments on attrition status, after controlling for stratification. It shows there is no statistically significant difference in attrition rates across treatment arms. Columns (2) and (4), show that the attriters are slightly older and less likely to be female, but columns (3) and (5) show that the reduced sample remains balanced on these two indicators. Column (6) shows that the attriters did not perform significantly better or worse at the baseline reading tests.

3.2.2 Survey data and document inspection

The teacher survey contained questions on basic demographics (education, gender, age, home language), teaching experience, curriculum knowledge, and teaching practice. For curriculum knowledge, we asked the number of times a week that the teacher performs the following activities: group-guided reading, spelling tests, phonics, shared reading, and creative writing. The prescribed frequency of performing these activities is stipulated in the government curriculum and also reflected in the lesson plans. Performing these activities at the appropriate frequency is thus a measure of knowledge and mastery of the curriculum, as well as fidelity to the scripted lesson plans.

The questions on teaching practice covered important pupil-teacher interactions that flow from group-guided reading: whether teachers ask pupils to read out loud, provide one-on-one assessment, and sort reading groups by ability. Finally, the teacher survey also included a voluntary comprehension test, which was completed by 75, 89, and 98 per cent of teachers who completed the teacher survey at baseline, midline and endline respectively.

In the enldine, we have teacher survey data for 275 teachers in 175 schools. As a result, for 81 percent of the 2,951 pupils assessed at endline, we also have data on their teacher.[12] In column (8) in Table A.1 we regress treatment assignment dummies on an indicator for whether a pupil's teacher also completed the teacher survey. We see that teacher non-response was random across treatment arms.

We also conducted classroom and document inspection for the surveyed teachers. Fieldworkers counted the number of days that writing exercises were completed in the exercise book, and the number of pages completed in the government workbook.[13] To minimize risk of bias due to strategic selection of exercise and workbooks, the teacher was asked to provide books of one of the most proficient pupils in his/her class. Furthermore, fieldworkers indicated if that teacher has a list for the reading groups, and rated on a 4-point Likert scale the sufficiency and quality of the following print material: a reading corner (box library), graded readers, Setswana posters, and flashcards.

[12] We cannot tell what proportion of teachers did not respond, because children are randomly drawn at a school level, so we do not know how many teachers pupils with missing teacher data would have matched with.

[13] To reduce data capture error, we asked the fieldworker to only count pages completed for three specific days. We chose three days that should have been covered by teachers by the end of the year, regardless of their choice of sequencing.

The school principal survey includes basic demographic questions, questions on school policies, school location, school access to resources, and a rough estimate of parent characteristics: the language spoken most commonly in the community, and highest overall qualification of the majority of parents.

3.2.3 Lesson observations

To gain a better understanding of how teaching practice changed in the classroom, we also conducted detailed lesson observations in October 2016 in a stratified random subset of 60 schools— 20 schools per treatment arm. We stratified by school performance in order to assure representation across the distribution of school performance. We also over-sampled urban schools, where the impacts of the programs were largest at midline.[14] An expert on early-grade reading developed the classroom observation instrument, in close consultation with ClassAct and the evaluation team.

The instrument covered teaching and classroom activities that we expect to be influenced by the program. For example, the fieldworkers were required to record the number of pupils who read or handle books; the number of pupils who practice the different building blocks of reading (e.g. vocabulary development, phonemic awareness, word/letter recognition, reading sentences or extended texts); and how reading is practiced in the classroom (e.g. read individually or in a group; read silently or loudly). The instrument also captured student-teacher interactions related to group-guided reading: Whether reading groups are grouped by ability, how frequently pupils receive individual feedback from the teacher, and how frequently pupils are individually assessed. These final set of indicators mirror the questions that were asked in the teacher survey.

The instrument was very detailed, but unlike other lesson observation instruments, did not require the fieldworkers to record time devoted to different activities. Rather, questions related to frequency of different activities were generally coded on a Likert scale.[15]

Since it was a detailed and comprehensive instrument, we decided to limit ourselves to six qualified fieldworkers, all of whom were proficient in Setswana and had at least a bachelors degree in teaching early-grade reading. To further assure consistency across fieldworkers, the project manager visited at least one school with each of the fieldworkers at the start of the data collection; and data quality checks were conducted after two days of data collection.

After the completion of the lesson observations, the fieldworkers also asked some questions about the type of teaching support they receive the past year. These were open-ended questions, which allowed us to code whenever a teacher mentions receiving training or coaching from Class Act, or using its graded readers or scripted lesson plans.

3.2.4 Administrative data

We further complemented these survey measures with 2011 census data and results from the 2014 Annual National Assessment (ANA). From the 2011 census, we constructed a community wealth index derived from several questions about household possessions, and we also calculated the proportion of 13 to 18

[14] In particular, we randomly drew schools from each treatment group in the following manner: (i) six urban schools; (ii) five schools in the top tercile and five schools in the bottom tercile in terms of average performance across both baseline and midline; (iii) four schools in the top tercile in terms largest improvement between baseline and midline.

[15] For example, when coding frequency of different types of reading activities, the fieldworkers recorded: never, sometimes, mostly, and always.

year-olds in the community that are attending an educational institution. We also have have data on each school's quintile in terms of socio-economic status, as coded by government.

3.2.5 Aggregation of indicators

In order to minimize the risk of over-rejection of the null hypotheses due to multiple different indicators, we aggregated data in the following ways. First, for own main outcome measure of success —reading proficiency— we combined all the sub-tasks into one aggregate score using principal components. We did this separately for each round of assessment. This score was then standardized across the whole sample to have a mean of zero and standard deviation of one. The treatment impact on the aggregate score can thus be interpreted in terms of standard deviations.

Furthermore, we grouped the potential mediating factors of changed teaching practice and classroom environment into five broad categories that are theoretically distinct inputs into learning to read: (i) access to reading material in the classroom; (ii) adherence to the teaching routine as prescribed in the curriculum; (iii) breadth of reading opportunities in the classroom; (iv) teacher-pupil interactions related to group-guided reading; and (v) pupils' use of reading material in the classroom. For each category we created a mean index, using the method proposed by Kling, Liebman and Katz (2007)

[TO DO: There was a slight modification of the grouping of mediating variables as originally specified in the pre-analysis plan. We believe this proposed grouping is more theoretically-motivated and a better reflection of the objectives of the program. In the appendix we present results as they were grouped in the pre-analysis plan.]

3.3 Balance and descriptive statistics

Table 1 shows balance and basic descriptive statistics of our evaluation sample. Each row represents a separate regression of the baseline variable on treatment assignments and strata dummies, clustering standard errors at the school level. The first column indicates the mean in the control. Columns (2) and (4) indicate the coefficient on dummies for the Training and Coaches interventions respectively.

Our sample of schools come predominantly from poor communities: 46.3 per cent of schools are in bottom quintile in terms of socio-economic status, and 85 per cent are from rural areas. In only 44 per cent of schools do the majority of parents have a high school degree or higher. In almost all schools the main language spoken in the community is Setswana. A sizable fraction of classrooms ended up being multi-grade classrooms (12.7 and 6.2 of grade one and grade two classrooms respectively). We were thus not perfectly able to identify and exclude *ex ante* all schools that do multi-grade teaching. Since the program was not intended to work in multi-grade settings, we will also report impacts on a sample that excludes these classrooms as well. Both the grade one and grade two teachers are mostly female and are educated: 85 and 95 per cent of the grade one and two teachers respectively have a degree or diploma. Nonetheless, reading comprehension levels are low: The average score for the simple comprehension test is 62 and 66 per cent for the grade one and two teachers respectively. We observe slight imbalance on baseline pupil reading proficiency and the school community's socio-economic status for the Training treatment arm. We control for all these variables in the main regression specification.

Table A.2. compares the sample where we conducted the lesson observations with the full evaluation sample. In each column we regress another independent variable on a dummy variable indicating if the pupil/school is in the sample where we conducted the lesson observation or not. In columns (1) to (4) the data is at the individual level; in column (5) the data is at the school level. In column (1) the dependent variable is midline reading proficiency, but we also include the full set of controls used in the main analysis (equation 1, below). A significant coefficient could thus interpreted as the 'value-added', over an above the average learning trajectory of a pupil. Columns (1) to (4) in table A.2. show that there is no statistically significant difference between schools where we conducted the lesson observations and the rest our evaluation sample, both in terms of pupil reading proficiency evaluated at baseline, midline and endline, and a value-added measure between baseline and endline. As expected given our sampling strategy, a far higher proportion of schools where we conducted lesson observations are urban: 36.7 per cent, compared to 20 per cent in our overall sample.

Figures A.2 to A.8 in the Appendix show the distribution of all the baseline sub-tasks, by treatment group. As to be expected we observe some floor effects for letter recognition, word recognition and comprehension test; and a ceiling effect for the vocabulary test. There is a good spread in the working memory tests. Encouragingly, the aggregate score shows good variation and no strong ceiling or floor effect.

3.4 Empirical Strategy

Our main estimating equation is:

$$y_{isb1} = \beta_0 + \beta_1 T_s + X'_{isb0}\Gamma + \rho_b + \varepsilon_{isb1}, \tag{1}$$

where y_{isb1} is the endline (end of second year) aggregate score of reading proficiency for pupil i in school s and strata b, $T_s \in (\text{Training}, \text{Coaching})$ is the relevant treatment dummy, ρ_b refers to strata fixed effects, X_{isb0} is a vector of baseline controls, and ε_{isb1} is the error term clustered at the school level. To estimate the respective impacts of the two interventions, we restrict the sample to the control schools and the schools from the relevant treatment group.

We control separately for each domain of reading proficiency collected at baseline: vocabulary, letter recognition, working memory, phonological awareness, word recognition, words read, and sentence comprehension. We control for each domain separately in order to increase statistical power. To further increase statistical power and account for any incidental differences that may exist between treatment groups, we control for individual and community-level characteristics which are highly correlated with y_{isb1} or were imbalanced at baseline.[16] Where data is missing for some observations for the control variables were imputed missing values[17] and added a dummy indicating missingness as a control.

When testing for heterogeneous treatment impacts, we estimate the following:

$$y_{isb1} = \beta_0 + \beta_1 T_s + \beta_2 (T \times \sigma)_s + X'_{isb0}\Gamma + \rho_b + \varepsilon_{isb1}, \tag{2}$$

[16] The additional controls include: pupil gender, pupils' parents' education, district dummy (schools were randomly spread across two districts), performance in the most recent standardized Annual National Assessments (ANA), a community-level wealth index, and average secondary school attendance rate in the community surrounding the school.

[17] For categorical variables, we assigned missing values to zero; for continuous variables we assigned missing observations to equal the sample mean.

where σ is a dummy variable referring to the sub-group of interest and is now also included in the vector of baseline controls. The coefficients on the interaction term, $(T \times \sigma)_s$, will thus show the difference in treatment impact across the sub-group of interest.

4 Results

4.1 Quality of implementation

As a first step in our analysis, we examine the quality of implementation. Panels (a) to (c) in Figure 1 and rows (1) to (3) in Table 2 show results from teacher questionnaire administered to all teachers in the evaluation sample. Panels (d) to (f) in Figure 1 and rows (4) to (6) in Table 2 show results from the in-depth teacher survey conducted in a sub-set of 60 schools.

We see that that the program was well-implemented: 97 and 94 per cent of teachers in the Training and Coaching arms respectively state that they have received in-service training on teaching Setswana as a home language the past year. The support was also generally well-received: 45 and 66 per cent in the Training and Coaching arms respectively state they received very good support in teaching Setswana, relative to 17 per cent in the Control.[18] Moreover, results from the sample of teachers interviewed during the lesson observations reveal that exposure to the program was high: 90% and 80% of the the teachers in the Coaching and Training arms respectively state to use the Class Act scripted lesson plans; 90 and 95 per cent respectively use the program's graded readers; and 80% of teachers in the Coaching arm reported that they were visited by the program's reading coach the past year. It also seems that there was some contagion where one school in the control reports to have received a visit from a reading coach and also uses Class Act's scripted lesson plans. This school remains in our sample of control schools.

4.2 Impacts on learning

Next we turn to the mean impacts of the programs on pupil reading proficiency. Figure 2 and the first column in Table 3 show the main regression results on the full sample, estimated using equation 1. Coaching (Panel A) has an estimated impact of 0.25 standard deviations on aggregate reading proficiency, compared to a statistically insignificant impact of 0.11 standard deviations for Training (Panel B). In both cases the impacts are larger when we exclude pupils in multi-grade classrooms: 0.18 and 0.3 standard deviations respectively. This is not surprising, since the program was never expected to be effective in multi-grade schools. Moreover, panel (c) in Figure 2 shows that the impacts are even larger when we exclude repeaters. These are pupils who had shorter exposure to the program, because they were not taught by the treated teachers in the second year. As an additional test, we also conduct randomization inference tests on the mean impacts. The p-value for the one-sided test of the sharp null hypothesis is 0.0554 and .001 for the Training and Coaching arms respectively.

Columns (3) to (9) in Table 3 further unpacks the results, looking separately at each domain of reading proficiency that constitutes the aggregate score. For sake of comparability, we also standardize these measures to have a mean zero and standard deviation of one. It is encouraging that the magnitude of the impact is roughly the same across all the domains of reading proficiency in the Coaching arm.

[18] Although interestingly teachers in the Coaching arm are more likely to state that they received too much support.

Even though the impact is slightly smaller in the case of writing (0.15 standard deviations), it remains statistically significant. For the Training arm, we only see statistical significance for phonemic awareness and non-word decoding. As discussed in Section 2, phonemic awareness and decoding form the basis for acquisition of reading proficiency. It is therefore encouraging that in the Training arm we see development in the fundamentals, since gains in other domains could follow. However, these results might also imply that the Training was not able to shift the more difficult teaching practices that are required for further development in reading proficiency: it is plausibly easier to teach phonemic awareness, compared to paragraph reading, since it is generally taught using whole-class shared reading rather than group-guided reading.

Rows (10) and (11) show impacts on additional indicators of learning that do not constitute the aggregate reading proficiency score. We asked simple questions to test proficiency in mathematics and English, because we were concerned of possible crowding out of other teaching activity. There is no reduction in mathematical ability. Moreover, we actually observe a statistically significant *improvement* in reading in English for the Coaching arm. Perhaps this is because learning to read in one's mother tongue makes it easier to learn to read in other language; or perhaps it is because the English teachers learnt new teaching techniques from the Setswana teachers.

4.3 Which schools to target?

As government or donors consider whether or not to expand these programs into new schools, an important consideration is any variation in effect sizes along school characteristics that are easily observable *ex ante*. This is because it is unlikely that the program can be simultaneously implemented in all poor schools in the country without substantially reducing the quality. Any scale-up will therefore inevitably be targeted at certain schools.

Panels (a) and (b) in Figure 3 show the heterogeneous treatment impacts for two school characteristics that are easily identified using administrative data: school location and socioeconomic status.[19] Table 4 shows the regression results associated with Figure 3, estimated using equation 2. It is clear from both panels (a) and (b) that there is dramatic geographic variation in program effectiveness. Pupil learning improved substantially in both the Coaching and Training treatment arms in urban schools, by 0.75 and 0.42 standard deviations respectively. In contrast, the impacts in rural schools are far smaller —0.06 and 0.1 standard deviations respectively— and we cannot reject the null hypothesis that there is no impact in these schools. Similarly, the impact is far larger in schools situated in relatively richer communities —0.31 and 0.36 standard deviations respectively— compared to schools in communities in the bottom quintile in terms of socioeconomic status. Table 4 further shows that the differences in effect size (that is, the coefficients on the interaction term) are statistically significant for both treatment arms and both moderating variables. Rows (3) and (6) in table 4 shows that the large rural negative interaction effects remain statistically significant if we drop repeaters and multi-grade classrooms from the sample: this trend is therefore not driven by the fact that there might be more multi-grade classrooms and/or grade repeaters in rural schools. This dramatic treatment heterogeneity in effect sizes between urban and rural schools raises the obvious question about why both programs are ineffective in rural areas, which will be

[19] Recall from section 3 that government classifies schools into quintiles, based on the socio-economic status of the community surrounding the school. Our study sample only consists of schools in the bottom three quintiles.

further examined in a different paper.[20]

Since there are only 36 urban schools in our sample (20% of our sample), there is some concern that the large impacts in urban schools are driven by a few outliers. As a further robustness check, we perform a jackknife resampling technique, running the interaction regression multiple times and each time dropping a different school. Table A.3. shows the descriptive statistics for the stored regression coefficients and p-values from the 180 regressions. The estimated impact in urban schools ranges between 0.68 and 0.84 standard deviations for the Coaching arm, and ranges between 0.33 and 0.47 standard deviations in the Training arm. Moreover, the coefficient on the interaction term for the Coaching arm is always significant ($p = 0.001$). The coefficient on the interaction term for Training is marginally insignificant for one iteration ($p = 0.108$). Nonetheless, the magnitude of the range of possible effect sizes in urban schools remains large.

Clearly, it will be most cost-effective if government targets urban schools. At first glance, this suggests a stark equity trade-off: the program is most effective in urban areas, where the communities are on average more affluent. However, the trade-off is not that stark in the South African context. Recall that the evaluation sample is already drawn from the bottom three quintiles of schools in terms of socioeconomic status.

Table A.4. shows that there is limited differences between urban and rural schools in terms of pupils' baseline reading proficiency, improvement in reading proficiency, and socio-economic status. Each column represents a separate regression on a dummy variable indicating if the school is rural or not.[21] Columns (1) and (2) show that pupils' reading proficiency when they enter school is no better in urban compared rural schools, nor is there a statistically significant difference in progress in the control schools over the two years of the program. Columns (3) and (4) show that parents in rural schools are less likely to have completed secondary school, and urban schools are slightly more likely to be in the bottom quintile in terms of socio-economic status. Nonetheless, the rates remain low in our sample of urban schools: only 36% of the pupils' parents had completed secondary school, and 36% of the schools are in the bottom quintile in terms of socio-economic status.

[Also include class size? And state that other teacher and school characteristics do not matter?]

4.4 Cost-effectiveness analysis

The results from any cost-effectiveness analysis (CEA) —i.e. calculating which program has the largest impact per dollar spent— depends clearly on the objectives of the social planner: the types of schools where it plans to scale up the program, the outcome indicator of interest, the welfare weight it places on different segments of the population, and the degree of risk it is willing to take. In this section we conduct CEA under different scenarios. We conclude that in all cases Coaching is more cost-effective than Training.

[20] In brief, we find that the treatment heterogeneity can be partly attributed to lower dosage in the rural areas: schools received fewer visits by coaches and were less likely to use the scripted lesson plans and graded readers. However, the difference in dosage is too small to explain the widely disparate impacts. Moreover, none of the observed teacher, school, and pupil characteristics can fully explain the difference.

[21] The regressions in columns (1) to (3) are estimated at the pupil level; the regression in column (4) is estimated at the school level. The regression in column (2) is restricted to the sample of pupils in control schools and also includes all the controls used in equation 1. The reported coefficient in column (2) can therefore be interpreted as "value-added": any advantage that rural schools face in terms of the speed of reading acquisition.

For cost estimates, we use the program budget for the third year of implementation. We choose the third year, since this is at a point where a lot of the set-up challenges have been resolved and fixed costs have been paid (all the materials have already been developed). One would therefore not expect the per-pupil cost to be much different when the program is scaled up to more schools. Based on these costs, the per pupil cost of the Coaching and Training programs are 48 USD and 62 USD respectively.[22]

For the simplest possible CEA we consider the whole evaluation sample and assume that the social planner is risk-neutral and has a constant marginal value to increasing the aggregate score. In this case Coaching is clearly more cost effective with a 0.41 standard deviation increase per 100 USD spent per pupil, compared to 0.23 increase in the case of Training. If we conduct cost-effectiveness analysis for the sub-sample of urban schools, where the program is most likely to be scaled up first, then Coaching remains more cost-effective with an estimated impact of 1.22 standard deviation increase per 100 USD spent per pupil, compared to 0.86 in the Training arm.

It is inherently difficult to to place value on an aggregated unit-free index, so in order to provide a more intuitive value for the CEA we consider a social planner who cares only about the number of pupils who meet a minimum level of literacy. For this we construct a dummy variable indicating whether a child passed the comprehension test or not.[23] With this indicator the Coaching arm over four times more cost-effective, with a 15.9 percentage point increase in the probability of reading with comprehension per 100 USD spent on a pupil, compared to a 3.3 percentage points increase in the Training arm.[24] In urban areas, Coaching remains more cost-effective at improving comprehension, with a 42.3 percentage point increase per 100 USD spent per pupil, compared to 18.3 in the case of Training.[25].

None of the above calculations consider risk and uncertainly, however. Given the large standard errors in effect sizes, we cannot with any large confidence conclude that Coaching is more cost-effective than Training. Nonetheless, incorporating any degree of risk aversion by a social planner would strengthen the case for Coaching. For example, consider a social planner that has a max-min objective function that only values the lower bound of a 95% confidence interval. If the social planner wishes to scale up in all poor schools, then Training will never be chosen because we cannot reject the null that it has zero impact. When restricting the sample to urban schools, Coaching is far more cost-effective, with a 0.73 SD increase per 100 USD, compared to 0.13 SD increase for Training.[26]

4.5 Changing teaching practice

In this section we further investigate underlying mechanisms by measuring how the learning environment, teaching practice, and classroom activities changed as a result of the program. For this purpose we draw from three different data-sources: the teacher survey administered in full evaluation sample of teachers,

[22] Total costs for implementing the program in 50 schools are 179, 853 USD and 230, 860 USD in the Coaching and Training arms respectively. Given an average size of 74.6 of pupils per school at the start of the program, this surmounts to per-pupil costs of 48 USD and 62 USD respectively.

[23] Reading comprehension is the ultimate goal of literacy development. However, it is an ambitious goal for grade 2 pupils and is arguably an unfair comparison, because Training has the lowest impact on comprehension, but statistically significant impacts on phonological awareness and letter recognition. Students who have only mastered these foundational skills could conceivably eventually progress into mastering literacy at later grades.

[24] We reach far stronger conclusions compared to using the aggregate reading proficiency score, because the average 0.112 standard deviation increase in the Training arm is not sufficient to pull pupils over the threshold of basic comprehension.

[25] p.point increase of 8.8 and 26.2 respectively

[26] Lower bound impacts are 0.45 and 0.06 SD for Coaching and Training respectively.

classroom and document inspection conducted in the same sample, and lesson observations conducted in a stratified random sub-set of 60 schools.

Two main results are worth emphasizing. First, even though there is no large difference in access to graded readers, the lesson observations reveal that far more pupils are reading the graded reading booklets in the program schools. This increase is substantially larger for teachers who received Coaching relative to teachers who received Training. Second, even though we find no change in the probability that pupils practicing reading in the classroom, there is a big difference in *how* they practice reading: Teachers in both Training and Coaching arms are more likely to enact group-guided reading, resulting in more students receiving more individual attention from teachers. Again, this impact is far larger for teachers who received Coaching relative to Training. These results suggest that there are some teaching practices that are difficult to enact and require additional coaching to be effective. They also reveal an important interaction between resources and teaching practice: the graded reading booklets are only useful if teachers have developed the skills to use them effectively in the classroom.

As discussed in section 3, we grouped the potential mediating factors into five broad categories: (i) access to reading material in the classroom; (ii) adherence to the teaching routine as prescribed in the curriculum; (iii) breadth of reading opportunities in the classroom; (iv) teacher-pupil interactions related to group-guided reading; and (v) pupils' use of reading material in the classroom. For each category we aggregated the indicators into a mean index, using the methodology proposed by Kling et al (2004). We also report results on each individual indicator. The first two categories —access to reading material and adherence to the teaching routine— provides an indication of at least superficial fidelity to the program. The subsequent two categories look at actual teaching activity in the classroom and assesses the enactment of different components of group-guided reading, an integral yet technically difficult component of the curriculum. The final category captures what is arguably one of the most important requirements for learning to read: opportunities for pupils to individually practice reading text. The regression results are reported in Tables 5 to 7. In all specifications we include stratification fixed effects and cluster our standard errors at the school level, where necessary.[27]

(i) Access to reading material in the classroom. Row (1) in Table 5 shows that there was a large and statistical significant improvement in overall access to reading material in the classroom: a 0.465 and 0.41 standard deviation increase for the Kling index in the Training and Coaching arms respectively. Rows (2) to (5) show results for indicators that constitute the mean index. For ease of interpretation we have converted the indicators from a 4-point categorical to binary variables.[28] There is a substantial increase in the probability that a classroom contains a well-stocked reading corner (a 25 and 26 percentage point increase in the Teaching and Coaching arms respectively), and exhibits a sufficient number of quality Setswana posters (25 and 21 percentage point increase respectively) and flash cards (a 18 and 17 per cent increase respectively) on the classroom wall. The magnitude of the impact is remarkably similar for both treatments. Notably, there is no impact on the probability that every pupil in the classroom has access to graded readers.

[27] We only observed one teacher per school in the lesson observations, so there is no need to cluster our standard errors at the school level. But we surveyed all the grade 2 teachers in each school, often more than one teacher per school.

[28] Results on statistical significance is the same when running an ordered probit model on the original ordinal variables. The reported mean index is constructed using the comprehensive 4-point indicators.

(ii) Fidelity to teaching routine Next we test for the hypothesis that teachers are more likely to follow the routine specified in the scripted lesson plans. In the teacher survey, we asked teachers how frequently they perform different types of teaching activities on a weekly basis: group-guided reading, spelling tests, phonics, shared reading, and creative writing.[29] Recall that the frequencies of doing these activities are clearly stipulated in the government curriculum, so in principle the teachers in the control should be performing them at the same frequency. We find teachers in both Training and Coaching schools are more likely to perform each activity at the appropriate level of frequency, relative to the control. Column (7) in row 14 shows that this impact is significantly higher for Coaches relative to Training. It is important to note that the treated teachers are not stating that they are more likely to perform *all* activities. They are more likely to perform activities that are required to be performed on a daily basis —group-guided reading and phonics— but state they are less likely to perform the activity that should only take place on a weekly basis— correcting spelling. These results can therefore not be attributed to pure experimenter demand effect of over-reporting all teaching activities.

We have learnt that teachers who received the scripted lesson plans claim to be more likely to follow the right routine, and as a result state they are more likely to teach phonics and facilitate group-guided reading in the classroom. Next, we deeper unpack whether they are more likely to perform these activities in the classroom.

(iii) Practicing reading and phonics Results from columns (1) to (7) in table 6 show that pupils are no more likely to practice reading in the classroom because of the program, nor is there any evidence that teachers are more likely to teach phonics.[30] We see in columns (8) and (9) that pupils in both the Training and Coaching arms are more likely to read extended texts, but the mean index is not significant.

One possible reason why we see no change in the probability that a teacher teaches phonics or letter or word recognition, is that this is typically done through whole class reading (where children repeats after the teacher), which is a relatively easy technique to enact. As a result, many teachers in the control are already teaching it. Reading multiple sentences, in contrast, is more difficult to teach because the curriculum requires that is taught through group-guided reading. More generally, these questions do not indicate *how* the pupils are practicing reading. We turn to this below.

(iv) Group-guided reading Next we deeper unpack the type of teaching activities related to group-guided reading, an activity that teachers in both Training and Coaching arm report to perform more frequently. Recall that there are three important components of group-guided reading: individual opportunities to read out loud, individual assessment, and sorting reading groups by ability. We asked for each one of these indicators separately in the teacher questionnaire, and also measured these activities during the lesson observations.

Rows (1) to (5) in Table 6 show result from the teacher survey. There was an overall increase for both

[29] Options were: Less than once a week, once a week, 2-4 times a week, every day, twice a day.

[30] The fieldworkers were asked to record how many pupils in the classroom are involved with reading letters, words, sentences, or extended texts. The answers were recorded as 5-point Likert scale, ranging from none to all the pupils. They also recorded the extent to which teacher covers phonics on a 4-point Likert scale. As before we construct binary variables for ease of interpretation (equal to one, if at least some pupils are reading; and equal to one if the teacher teaches phonics at least some of the time). Results on statistical significance remain the same when running an ordered logit model on the ordinal variables.

treatment arms in the activities that relate to group-guided reading, with a consistently larger impact for Coaching relative to Training. First, as a confirmation of the self-reported increase in conducting group-guided reading, we find that teachers that the program teachers are more likely to provide a list of reading groups relative to the control (17 and 24.4 per cent in the Training and Coaching arms respectively), and this impact is significantly larger for teachers that received Coaching. We further find that teachers that received Coaching are far more likely than the Training and control teachers to state that they listen almost daily to pupils read out loud, and more likely to perform one-on-one reading assessment at least weekly. Teachers in both Training and Coaching state are more likely to state that they stream groups by ability.

The results from the teacher survey provide strong evidence that group-guided reading was far more likely to take place in both treatment arms, with the largest increase observed for teachers who received Coaching. Moreover, the larger change seems to come from individual attention, rather than streaming. However, these results are all self-reported. To test if these practices actually changed in the classroom, we next turn to results from the lesson observations.

Rows (6) to (11) in Table 6 shows that the results from the teacher survey on group-guided reading are broadly supported by the lesson observations: there is a large, statistically significant increase in the mean index of 0.73 and 0.86 standard deviations in the Training and Coaching groups respectively. When examining the different components of group-guided reading, we see that there is a large increase in the Coaching arm in the probability that the pupils are split into groups (55.5 percentage point increase), that pupils read aloud in groups (41 percentage point increase), and that the pupils read individually to the teacher (51.5 percentage point increase).[31] The impact for these three indicators is smaller for the Training arm, and not always statistically significant. However, we do not find strong evidence for any improvement in the probability of providing individual assessment and grouping by ability.[32] Note, that not *all* types of reading activities are more likely to take place. Rows (12) to (14) show that teachers are no more likely to perform whole-class reading, where the whole class reads aloud with the teacher. Teachers are also no more/less likely to read aloud with the pupils following silently. Whole-class reading is an easy activity to perform in the classroom, and almost all teachers in the control are already doing it.

Taken together we see strong evidence that there was an increase in group-guided reading in both treatment arms, with the largest change observed for teachers that received the Coaching. This coincided with more individual attention by the teacher and opportunities to read out loud in groups, but there is weaker evidence for any change in individual assessment and sorting by ability. The fact that these activities are more likely to take place in the Coaching arm suggests that group-guided reading is a pedagogical skill that requires the additional monitoring and feedback from coaches to fully develop. This is also suggestive evidence that these activities related to group-guided reading are at least part of

[31] The two latter indicators —reading aloud in groups and individual attention— are ordinal variables ranked from 1 to 4. For ease of interpretation we created a binary indicator for these two indicators, indicating if *any* activity took place. The mean index, however, is constructed using the ordinal variable, thus preserving all the information captured by fieldworkers. Results on statistical significance remain the same when running an ordered logit model on the ordinal variables.

[32] There is a small increase in the probability of providing individual assessment, which is statistically significant only in the Training arm. Teachers that received Coaching are 24.7 percentage more likely to have different reading groups assigned to different graded readers (compared to 9.2 percentage point increase for teachers that received Training), but the difference is not statistically significant.

the explanation for faster acquisition of reading proficiency in the Coaching arm relative to Training.

(v) Pupil use of reading material As a final measure, we also test if pupils have more individual opportunities to read text. During the lesson observations, fieldworkers were required to count how many pupils have an opportunity to hold books (excluding the government workbooks) and how many pupils read the graded reading booklets. Even though there was no difference in *access* to the graded readers between any treatment arm and the control, we see a substantial increase in *use* of reading material, especially in the number of children who have an opportunities to read. These results are reported in rows (15) to (17) in Table 7. Strikingly, in the control schools only one pupil in one school read a book, leading to an average of 0.05 pupils reading a book in the control. The average number of pupils who read increased by 2.3 and 5.1 in the Training and Coaching arms respectively. These is also a marked difference between the treatment arms: far more pupils in the Coaching arm handle books and read the graded readers.

These results reveal the important interaction between resources, teaching practice, and use of resources. Access to graded readers is high in all the evaluation arms, including the control. However the purpose of the graded readers is to provide individual opportunities to practice reading. Pupils are provided this opportunity during group-guided reading, an activity that teachers find challenging to implement. These resources therefore cannot be used without appropriate enactment of a new teaching method. Consequently, very few pupils are actually reading graded readers in the control schools.[33]

5 Conclusion

In this paper we report the endline results of a randomized evaluation of interventions aimed at improving early grade reading. Both interventions provide scripted lesson plans and additional supporting material such as graded readers and flash cards. They differ in the mode of delivery and cost of implementation. In one intervention (Training), teachers participate in two two-day training programs conducted at a central venue. In the other intervention (Coaching), teachers are visited at a monthly basis by specialist reading coaches who monitor their teaching and provide feedback.

We find that Coaching had a large and statistically significant impact on pupil reading proficiency, more than twice the size of the Training arm. We also find that the benefits of the program are unequally shared. Pupils from rural schools do not benefit from the program, but the impacts are immense in urban schools. Furthermore, the pupils who entered school with the largest shortfall in reading proficiency do not benefit at all.

The fact that the Training arm was ineffective (or at least, we were unable to detect an impact) is important to note for the design of in-service teacher training programs. Across the world, the most common form of government training is short training at a central venue. This program was implemented by a highly motivated non-governmental organization with strong incentives to demonstrate impact, and is based on strong pedagogical theory. The feedback from teachers was overwhelmingly positive. Yet, the impact remained modest. It turns out that changing pedagogical practice is difficult, and governments

[33] As further suggestive evidence, there is a strong correlation between number of pupils reading a graded reader and whether the teacher practiced group-guided reading or not, even after controlling for treatment assignment and stratification.

school should rather consider programs that provide ongoing pedagogical feedback and support, rather than once-off training.

6 Tables and Figures

Table 1. Descriptive and balance statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Control	*Training*		*Coaching*			
	Mean	Coef.	Std error	Coef.	Std error	Obs	R-squared
Pupil Characteristics							
Age	6.481	0.0781	(0.0520)	-0.0244	(0.0524)	3,523	0.011
Female	0.479	-0.0156	(0.0220)	-0.0120	(0.0207)	3,518	0.001
Reading proficiency	0.0380	-0.209*	(0.118)	0.0666	(0.146)	3,539	0.058
Grd 1 Teacher Characteristics							
Diploma or degree	0.852	-0.0489	(0.0500)	0.0774*	(0.0442)	294	0.043
Age	48.84	-1.565	(1.306)	2.044*	(1.232)	294	0.040
Female	0.982	0.0219	(0.0147)	-0.00848	(0.0272)	259	0.035
Class size	42.83	-1.082	(1.738)	-3.487**	(1.742)	291	0.139
Multi-grade	0.127	-0.0113	(0.0527)	0.0627	(0.0545)	292	0.090
Comprehension test	0.622	0.0163	(0.0326)	0.0157	(0.0315)	286	0.069
Grd 2 Teacher Characteristics							
Diploma or degree	0.947	0.0127	(0.0312)	0.0413	(0.0253)	271	0.021
Age	48.92	-1.566	(1.365)	-0.287	(1.217)	273	0.017
Female	1	-0.0138	(0.0134)	0.00001	(0.00234)	271	0.069
Class size	42.17	-1.993	(1.464)	-3.174**	(1.589)	271	0.131
Multi-grade	0.0619	0.00698	(0.0333)	0.00253	(0.0293)	271	0.242
Comprehension test	0.663	-0.0425	(0.0304)	-0.00419	(0.0326)	269	0.043
School characteristics							
Setswana most common	1	-0.0418	(0.0284)	-0.0216	(0.0213)	167	0.095
Majority parents - highschool	0.443	-0.106	(0.0871)	0.0341	(0.0823)	179	0.155
Rural	0.850	-0.0700	(0.0679)	-0.110	(0.0691)	180	0.177
Bottom quintile (SES)	0.463	0.0975*	(0.0520)	-0.0425	(0.0392)	180	0.757
Pass rate (ANA)	55.35	-1.184	(0.894)	-0.981	(0.917)	180	0.583
Kenneth district	0.212	-0.0125	(0.0705)	0.0875	(0.0771)	180	0.123

Notes: Each row indicates a separate regression on treatment dummies controlling for strata indicators. Column one shows the control mean, columns (2) and (4) the coefficient on the two treatment dummies. Standard errors (columns (3) and (5)) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table 2: Implementation

	(1)	(2)	(3)	(4)	(5)	(6)
	General support in teaching Setswana			*Exposure to Class Act (in-depth teach survey)*		
	Received training	Have graded reader	Very good support	Graded readers	Lesson plans	Reading coach
Training	0.179***	0.0300	0.287***	0.720***	0.732***	0.0167
	(0.0471)	(0.0272)	(0.0696)	(0.118)	(0.124)	(0.105)
Coaching	0.148***	0.0342	0.490***	0.677***	0.858***	0.792***
	(0.0514)	(0.0266)	(0.0637)	(0.117)	(0.0938)	(0.104)
Observations	274	263	272	60	60	60
R-squared	0.106	0.054	0.232	0.726	0.629	0.666
Control mean	0.793	0.956	0.167	0.150	0.100	0.0500

Notes: each column represents a separate regression, including strata fixed effects. Date from rows (1) to (3) come from the teacher questionnaire administered to all teachers in the evaluation sample. Data from rows (4) to (6) come from the in-depth teacher survey conducted in a sub-set of 60schools. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table 3. Main results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	`(11)
						Paragraph	Comprehe	Phon.			
	Aggregate Score		Letters	Words	Non-words	reading	nsion	awareness	Writing	Math	English
Panel A											
Coaching	0.252***	0.297***	0.200**	0.233***	0.265***	0.241***	0.233***	0.160**	0.153*	0.0542	0.221***
	(0.0792)	(0.0817)	(2.13)	(3.18)	(3.56)	(3.49)	(3.23)	(2.27)	(1.81)	(0.79)	(3.25)
Excl. multi-grade	N	Y	N	N	N	N	N	N	N	N	N
Observations	2,140	1,986	2140	2140	2140	2140	2140	2140	2140	2140	2140
R-squared	0.178	0.184	0.114	0.162	0.143	0.162	0.165	0.158	0.155	0.139	
Panel B											
Training	0.112	0.176**	0.0381	0.103	0.130*	0.123	0.0479	0.130**	0.103	0.0488	0.0924
	(0.0814)	(0.0843)	(0.43)	(1.27)	(1.68)	(1.62)	(0.65)	(2.02)	(1.29)	(0.71)	(1.19)
Excl. multi-grade	N	Y	N	N	N	N	N	N	N	N	N
Observations	2,121	2,001	2121	2121	2121	2121	2121	2121	2121	2121	2121
R-squared	0.170	0.177	0.134	0.168	0.148	0.170	0.157	0.173	0.165	0.158	

Notes: each column represents a separate regression, using equation (1). The top column indicates the outcome variables. In Panel A the sample is restricted to the Training and Control schools. In Panel B the sample is restricted to the Coaches and Control schools. In row (2) the sample is further restricted to schools that do not have multi-grade classrooms. Standard errors are in parentheses and clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table 4. Treatment impacts by location and socio-economic status

	(1)		(2)	(3)		(4)
	Training			*Coaching*		
Treat	0.416**	0.540***	0.312***	0.755***	0.802***	0.363***
	(0.178)	(0.205)	(0.111)	(0.153)	(0.183)	(0.0956)
Treat x Rural	-0.358*	-0.350*		-0.650***	-0.585***	
	(0.187)	(0.211)		(0.181)	(0.220)	
Treat x Quintile 1			-0.348**			-0.298**
			(0.144)			(0.146)
Excl. multi-grade classrooms and repeaters	N	Y	N	N	Y	N
Observations	2,121		2,121	2,140		2,140
R-squared	0.175		0.188	0.192		0.190

Notes: each column represents a separate regression, using equation (2). In columns (1) and (2) the treatment dummy is Training; in columns (3) and (4) the treatment dummy is Coaches. In columns (1) and (3) treatment is interacted with rural; in columns (2) and (4) treatment is interacted with quintile. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Access to print and adherence to teaching routine

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Control	*Training*		*Coaching*			p-value
								Training =
		mean	Coef.	Std. Error	Coef.	Std. Error	Obs	Coaches
Access to reading material								
(1)	*Kling index*	0	0.465***	(0.120)	0.410***	(0.114)	264	0.651
(2)	All have graded readers	0.416	0.114	(0.0921)	0.0327	(0.0904)	263	0.449
(3)	Reading corners	0.486	0.252***	(0.0854)	0.260***	(0.0806)	253	0.930
(4)	Setswana posters	0.316	0.249***	(0.0821)	0.206**	(0.0865)	263	0.651
(5)	Flash cards	0.752	0.177***	(0.0564)	0.166***	(0.0592)	263	0.828
Routine								
(6)	*Kling index*	0	0.300***	(0.0811)	0.497***	(0.0652)	276	0.0209
(7)	Group-guided reading	0.241	0.124*	(0.0738)	0.197***	(0.0674)	274	0.363
(8)	Spelling test	0.696	0.155**	(0.0627)	0.238***	(0.0509)	273	0.143
(8)	Phonics	0.491	-0.0708	(0.0745)	0.171**	(0.0720)	274	0.00195
(9)	Shared reading	0.422	0.183**	(0.0728)	0.171**	(0.0711)	274	0.872
(10)	Creative writing	0.310	0.301***	(0.0715)	0.383***	(0.0681)	274	0.286

Notes. Each row represents a separate regresion, including stratification fixed effects. Data is at the teacher level. Standard errors are clustered at the school level *** p<0.01, ** p<0.05, * p<0.1

Table 6 . Reading Frequency and Use of Reading Material as reported by lesson observations

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Control	*Training*		*Coaching*			Training = Coaching
		mean	Coef.	Std. error	Coef.	Std. error	Obs	P-value
Reading frequency (lesson observations)								
(1)	*Kling index*	0	0.0767	(0.149)	0.148	(0.150)	60	0.300
(2)	Phonics	0.684	0.148	(0.156)	0.135	(0.167)	59	0.927
(3)	Letters	0.625	-0.126	(0.185)	0.105	(0.174)	49	0.231
(4)	1-2 words	0.471	-0.0408	(0.176)	0.229	-0.227	44	0.378
(5)	3-10 words	0.667	-0.0582	(0.148)	0.0905	(0.129)	52	0.425
(6)	10+ words	0.133	0.0772	(0.151)	0.111	(0.170)	40	0.406
(7)	1-2 sentences	0.529	-0.269	(0.201)	-0.115	(0.214)	44	0.268
(8)	3-5 sentences	0.333	0.389**	(0.178)	0.441***	(0.161)	48	0.360
(9)	5+ sentences	0.188	0.352**	(0.173)	0.363**	(0.177)	49	0.330
(10)	Extended texts	0.579	0.0262	(0.181)	0.148	(0.182)	55	0.237

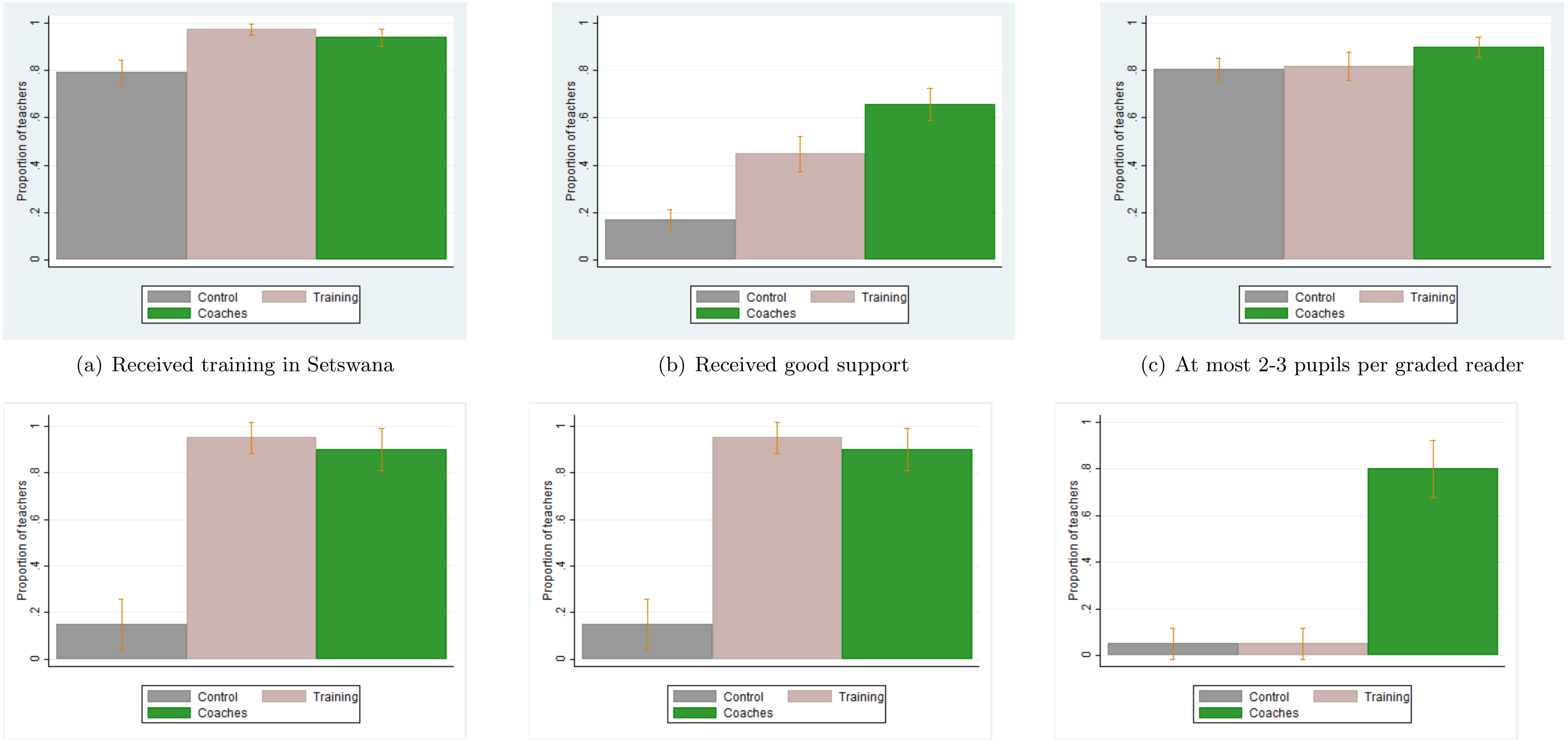
Notes. Each row represents a separate regresion, including stratification fixed effects. Data is at the teacher level, each teacher at a different school. *** p<0.01, ** p<0.05, * p<0.1

Table 7. Group-guided reading and use of reading material

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Control	*Training*		*Coaching*			Training = Coaching
		mean	Coef.	Std. error	Coef.	Std. error	Obs	P-value
Group-guided reading (questionnaire)								
(1)	*Kling index*	0	0.210**	(0.0880)	0.415***	(0.0772)	276	0.0124
(2)	Teacher can provide list of groups	0.430	0.168*	(0.0987)	0.344***	(0.0815)	232	0.0748
(3)	Listen to each pupil read out loud	0.578	0.0324	(0.0772)	0.237***	(0.0638)	273	0.00714
(4)	One-on-one reading assessment	0.655	0.0877	(0.0755)	0.161**	(0.0638)	274	0.296
(5)	Stream by ability	0.718	0.107*	(0.0579)	0.144**	(0.0580)	261	0.527
Group-guided reading (lesson observations)								
(6)	*Kling index*	0	0.725***	(0.243)	0.864***	(0.230)	60	0.538
(7)	Pupils split into groups	0.211	0.365**	(0.169)	0.555***	(0.160)	52	0.252
(8)	Pupils read aloud in groups	0.444	0.140	(0.194)	0.410**	(0.158)	54	0.102
(9)	Pupils read individually to teacher	0.176	0.334*	(0.186)	0.515***	(0.183)	51	0.317
(10)	Individual reading assessment	0.158	0.295*	(0.170)	0.125	(0.177)	55	0.340
(11)	Reading groups, different texts	0.105	0.0919	(0.133)	0.247	(0.161)	52	0.415
Whole class reading								
(12)	Teacher reads, class not following	0.222	-0.195	(0.128)	-0.0271	(0.146)	50	0.207
(13)	Teacher reads, class following silently.	0.550	-0.0889	(0.199)	0.0565	(0.215)	52	0.437
(14)	Whole class reads aloud with teacher	0.833	-0.0851	(0.183)	0.150	(0.127)	50	0.115
Use of reading material								
(15)	*Kling index*	0	4.859*	(2.551)	12.15***	(2.532)	60	0.004
(16)	No. learners handle books	1	0.717	(0.988)	2.542**	(1.001)	59	0.0145
(17)	No. learners read readers	0.0526	2.329**	(1.098)	5.093***	(1.067)	57	0.009

Notes. Each row represents a separate regresion, including stratification fixed effects. Data is at the teacher level. Data from rows (1) to (5) come from the teacher survey conducted in the full evaluation sample. Data from rows (6) to (17) come from lesson observations conducted in a sub-sample of 60 schools. *** p<0.01, ** p<0.05, * p<0.1

Figure 1: Implementation quality



(a) Received training in Setswana

(b) Received good support

(c) At most 2-3 pupils per graded reader

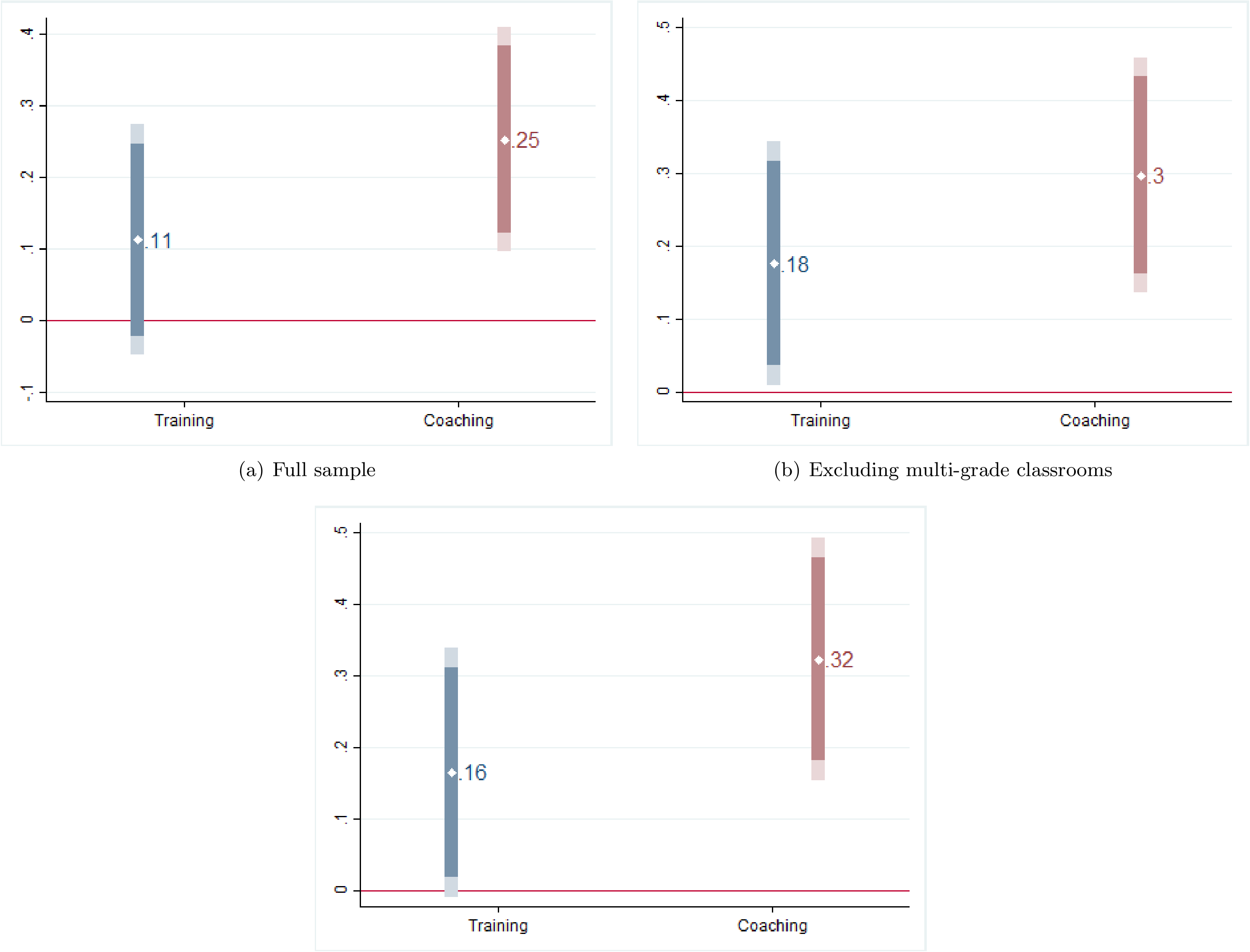
(d) Use ClassAct's graded readers

(e) Use ClassAct's scripted lesson plans

(f) Monitored by ClassAct reading coach

Note: Panels (a) to (c) are based on teacher questionnaires answered by 276 teachers (118, 76, and 82 teachers in the Control, Training and Coaching arms respectively) in 175 schools (77 schools in the Control, and 49 schools in each intervention arm). Panels (d) to (f) are based on in-depth teacher surveys conducted with a subset of 60 teachers, 20 teachers in each treatment arm

Figure 2: Mean impacts on learning

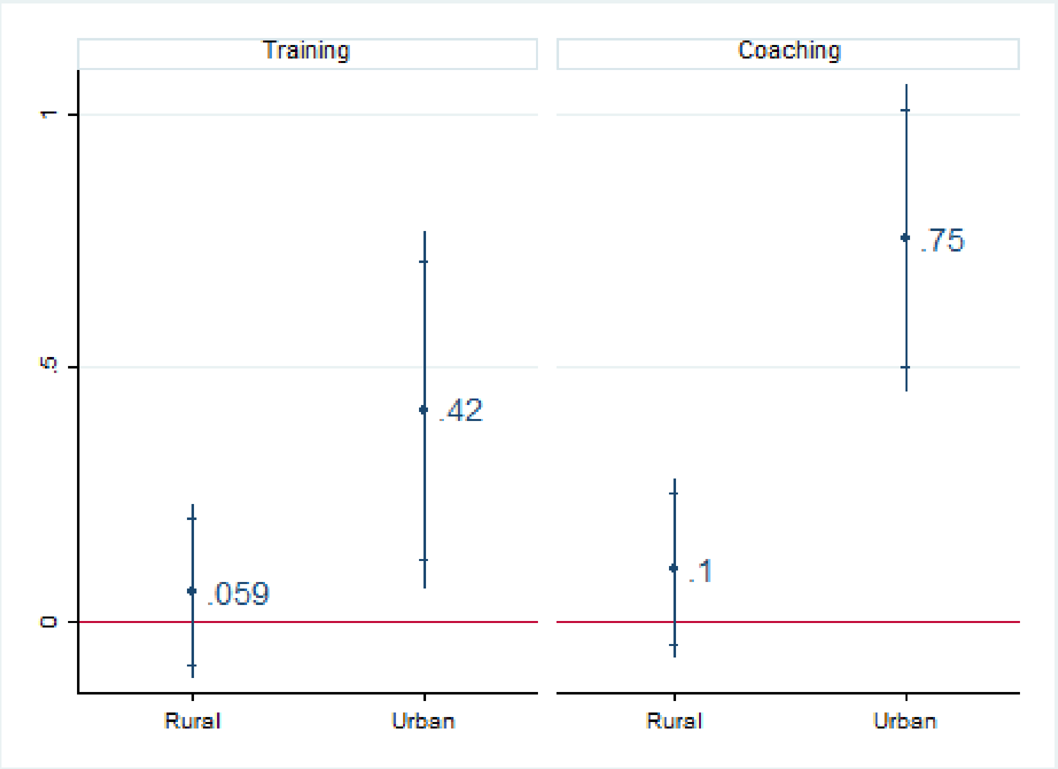


(a) Full sample

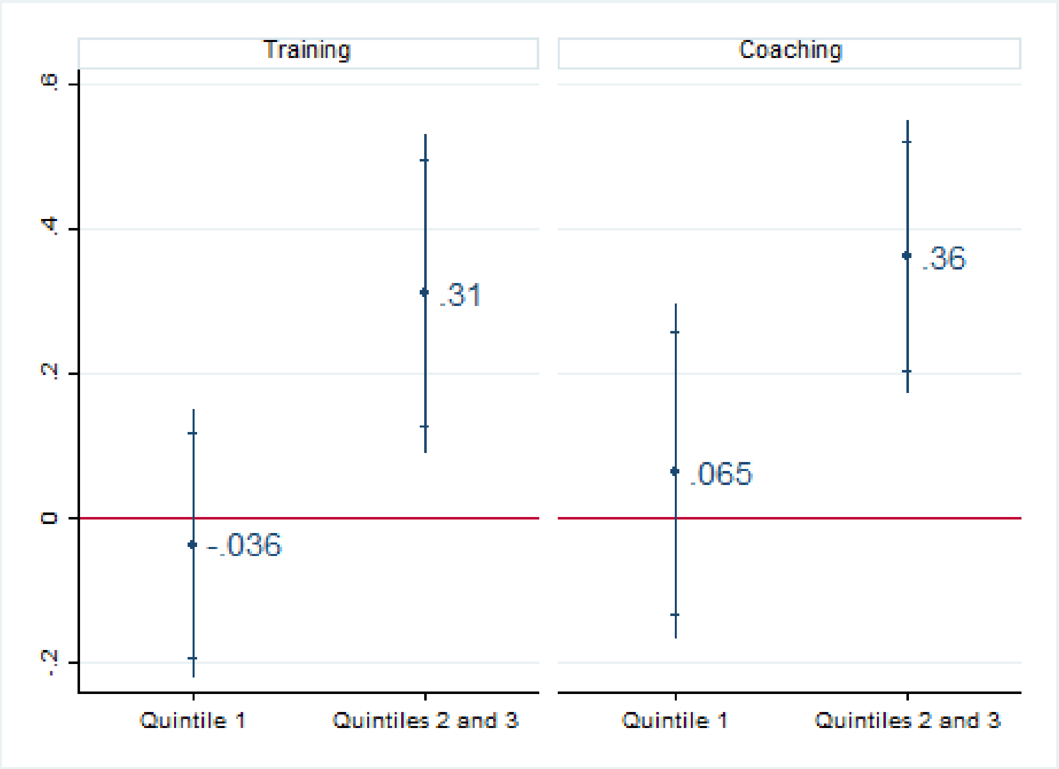
(b) Excluding multi-grade classrooms

(c) Excluding repeaters

Figure 3: Impacts by school location and socio-economic status



(a) Rural vs Urban schools



(b) Socio-economic status

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B. & Pianta, R. (2013), 'Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary', *School Psychology Review* **42**(1), 76.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y. & Schady, N. (2014), 'A helping hand? teacher quality and learning outcomes in kindergarten', *Banco Interamericano de Desarrollo, Washington, DC. Inédito* .

Banerjee, A. V., Cole, S., Duflo, E. & Linden, L. (2007), 'Remedying education: Evidence from two randomized experiments in india', *The Quarterly Journal of Economics* pp. 1235–1264.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014), 'Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood', *The American Economic Review* **104**(9), 2633–2679.

Cilliers, J., Kasirye, I., Leaver, C., Serneels, P. & Zeitlin, A. (2014), 'Pay for locally monitored teacher attendance?'.

De Ree, J., Muralidharan, K., Pradhan, M. & Rogers, H. (2015), Double for nothing? experimental evidence on the impact of an unconditional teacher salary increase on student performance in indonesia, Technical report, National Bureau of Economic Research.

Duflo, A. & Kiessel, J. (2014), 'Every child counts: Adapting and evaluating research results on remedial education across contexts.', *Society for Research on Educational Effectiveness* .

Duflo, E., Dupas, P. & Kremera, M. (2011), 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya', *The American Economic Review* **101**(5), 1739–1774.

Evans, D. K. & Popova, A. (2015), 'What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews', *World Bank Policy Research Working Paper* (7203).

Fryer, R. G. (2017), 'The production of human capital in developed countries: Evidence from 196 randomized field experimentsa', *Handbook of Economic Field Experiments* **2**, 95–322.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H. S., Doolittle, F. et al. (2008), 'The impact of two professional development interventions on early reading instruction and achievement. ncee 2008-4030.', *National Center for Education Evaluation and Regional Assistance* .

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P. et al. (2011), 'Middle school mathematics professional development impact study: Findings after the second year of implementation', *National Center for Education Evaluation and Regional Assistance* .

Glewwe, P., Ilias, N. & Kremer, M. (2010), 'Teacher incentives', *American Economic Journal: Applied Economics* **2**(3), 205–227.

Harris, D. N. & Sass, T. R. (2011), 'Teacher training, teacher quality and student achievement', *Journal of public economics* **95**(7), 798–812.

Jackson, C. K. & Makarin, A. (2016), Simplifying teaching: A field experiment with online" off-the-shelf" lessons, Technical report, National Bureau of Economic Research.

Jacob, A. & McGovern, K. (2015), 'The mirage: Confronting the hard truth about our quest for teacher development.', *TNTP* .

Jacob, B. A. & Lefgren, L. (2004), 'The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in chicago', *Journal of Human Resources* **39**(1), 50–79.

Kane, T. J. & Staiger, D. O. (2008), Estimating teacher impacts on student achievement: An experimental evaluation, Technical report, National Bureau of Economic Research.

Kane, T. J. & Staiger, D. O. (2012), 'Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. research paper', *Bill & Melinda Gates Foundation* .

Kennedy, M. M. (2016), 'How does professional development improve teaching?', *Review of Educational Research* **86**(4), 945–980.

Kerwin, J. T., Thornton, R. et al. (2015), 'Making the grade: Understanding what works for teaching literacy in rural uganda', *Unpublished manuscript. University of Illinois, Urbana, IL* .

Kling, J. R., Liebman, J. B. & Katz, L. F. (2007), 'Experimental analysis of neighborhood effects', *Econometrica* **75**(1), 83–119.

Lucas, A. M., McEwan, P. J., Ngware, M. & Oketch, M. (2014), 'Improving early-grade literacy in east africa: Experimental evidence from kenya and uganda', *Journal of Policy Analysis and Management* **33**(4), 950–976.

Muralidharan, K. & Sundararaman, V. (2009), Teacher performance pay: Experimental evidence from india, Technical report, National Bureau of Economic Research.

Muralidharan, K. & Sundararaman, V. (2013), Contract teachers: Experimental evidence from india, Technical report, National Bureau of Economic Research.

Neal, D. T., Wood, W. & Quinn, J. M. (2006), 'Habitsa repeat performance', *Current Directions in Psychological Science* **15**(4), 198–202.

Piper, B. & Korda, M. (2011), 'Egra plus: Liberia. program evaluation report.', *RTI International* .

Piper, B., Zuilkowski, S. S. & Mugenda, A. (2014), 'Improving reading outcomes in kenya: First-year effects of the primr initiative', *International Journal of Educational Development* **37**, 11–21.

Popova, A., Evans, D. K. & Arancibia, V. (2016), ‘Inside in-service teacher training: What works and how do we measure it?’.

Rivkin, S. G., Hanushek, E. A. & Kain, J. F. (2005), ‘Teachers, schools, and academic achievement’, *Econometrica* **73**(2), 417–458.

Appendix A Further tables and figures

Table A.1 Treatment Status Regressions on Attrition Status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Attrite	Age		Female		Reading proficiency		Teacher attrition
Training	0.00605		0.0870		-0.0307		-0.205*	-0.0239
	(0.0222)		(0.0530)		(0.0241)		(0.121)	(0.0363)
Coaching	-0.0136		-0.0246		-0.0139		0.0827	-0.0234
	(0.0183)		(0.0513)		(0.0230)		(0.151)	(0.0378)
Attrite		0.163***		-0.0357*		-0.0720		
		(0.0425)		(0.0215)		(0.0509)		
Observations	3,539	3,523	2,941	3,518	2,934	3,539	2,951	2,951
R-squared	0.010	0.007	0.012	0.001	0.003	0.001	0.061	0.013
Control mean	0.168							0.208

Table A.2 Comparing lesson observation schools with full sample

	(1)	(2)	(3)	(4)	(5)
	Pupil reading proficiency				*Location*
	Value-added	Endline	Midline	Baseline	Rural
In sample	0.0594	-0.00586	0.0200	-0.0284	-0.250***
	(0.0724)	(0.0814)	(0.0748)	(0.119)	(0.0692)
Observations	3,148	3,148	3,337	3,539	180
R-squared	0.001	0.000	0.000	0.000	0.087
Sample mean	0.0368	0.00873	0.0304	-0.0180	0.633

Notes: Each column represents a separate regression on a dummy variable indicating whether the pupil/school is in the sample where we conducted the lesson observation or not. In columns (1) to (4) the data is at the individual level; in column (5) the data is at the school level. In column (1) the dependent variable is midline reading proficiency, but we also include the full set of controls used in table 2.

Table A.3 Jacknife Resampling

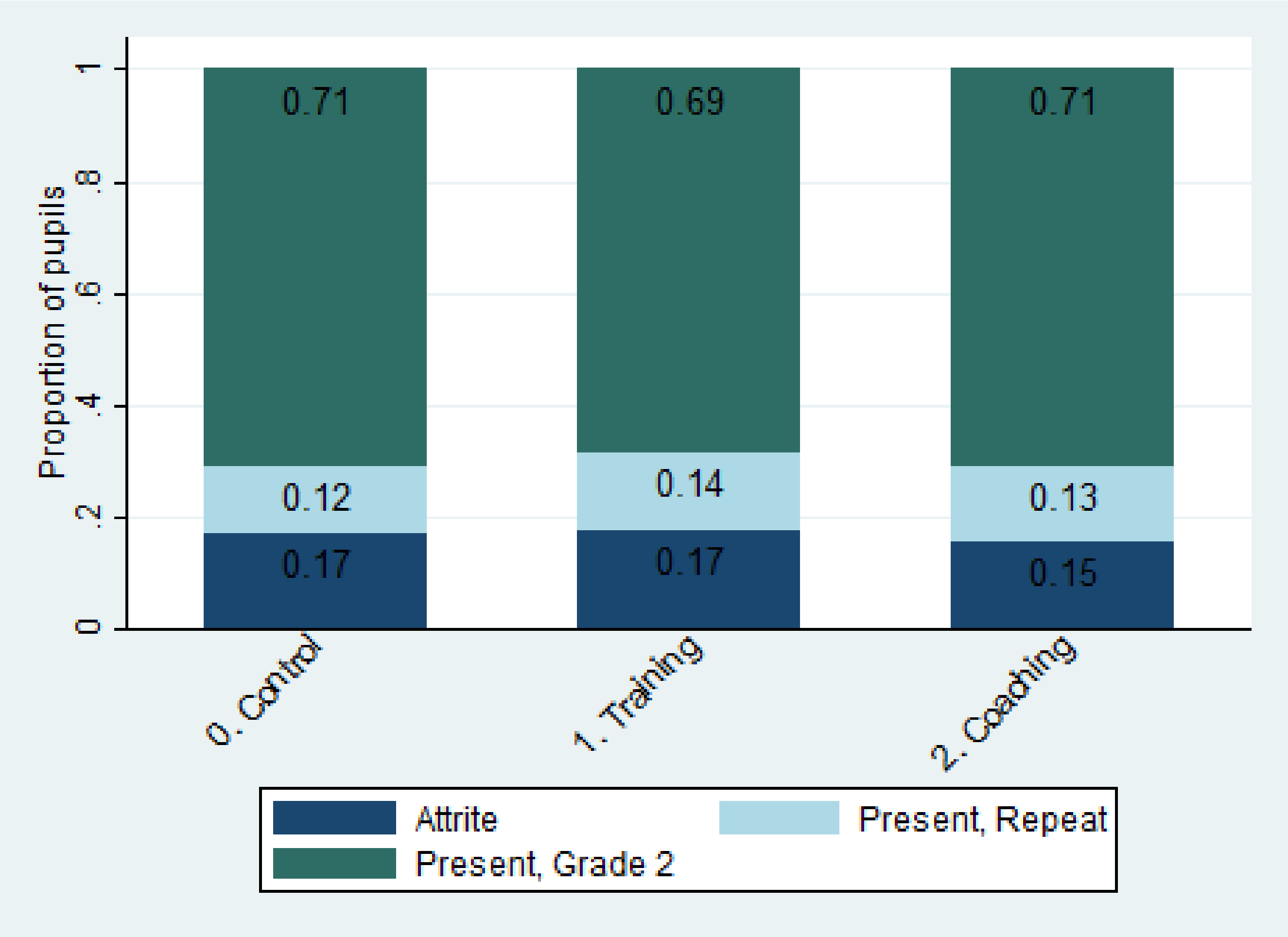
	(1)	(2)	(3)	(4)	(5)	(6)
	Coefficient			*P-Value*		
	mean	min	max	mean	min	max
Training	0.41	0.33	0.47	0.02	0.005	0.040
Coaching	0.76	0.68	0.84	<0.001	<0.001	<0.001
Rural x Training	-0.36	-0.42	-0.28	0.05	0.022	0.108
Rural x Coaches	-0.66	-0.74	-0.58	<0.001	<0.001	0.001

Table A.4 Differences between urban and rural schools

	(1) Baseline reading	(2) Value-added	(3) Parent completed highschool	(4) Quintile one school
Rural	-0.0825	0.149	-0.0601*	0.146
	(0.143)	(0.190)	(0.0309)	(0.0908)
Observations	3,539	1,310	3,118	180
Urban mean	0.0665	0.126	0.361	0.361
R-squared	0.001	0.199	0.003	0.014

Value added: regression in control schools with all main controlsused in regression.

Figure A.1: Attrition and repetition rates across treatment arms



Note: The figure shows the proportion of surveyed pupils by treatment group who: (i) were not present at endline for the reading assessment; (ii) were present, but are repeating grade one; (iii) were present and are in grade two

Figure A.2: Baseline distribution - picture comprehension test

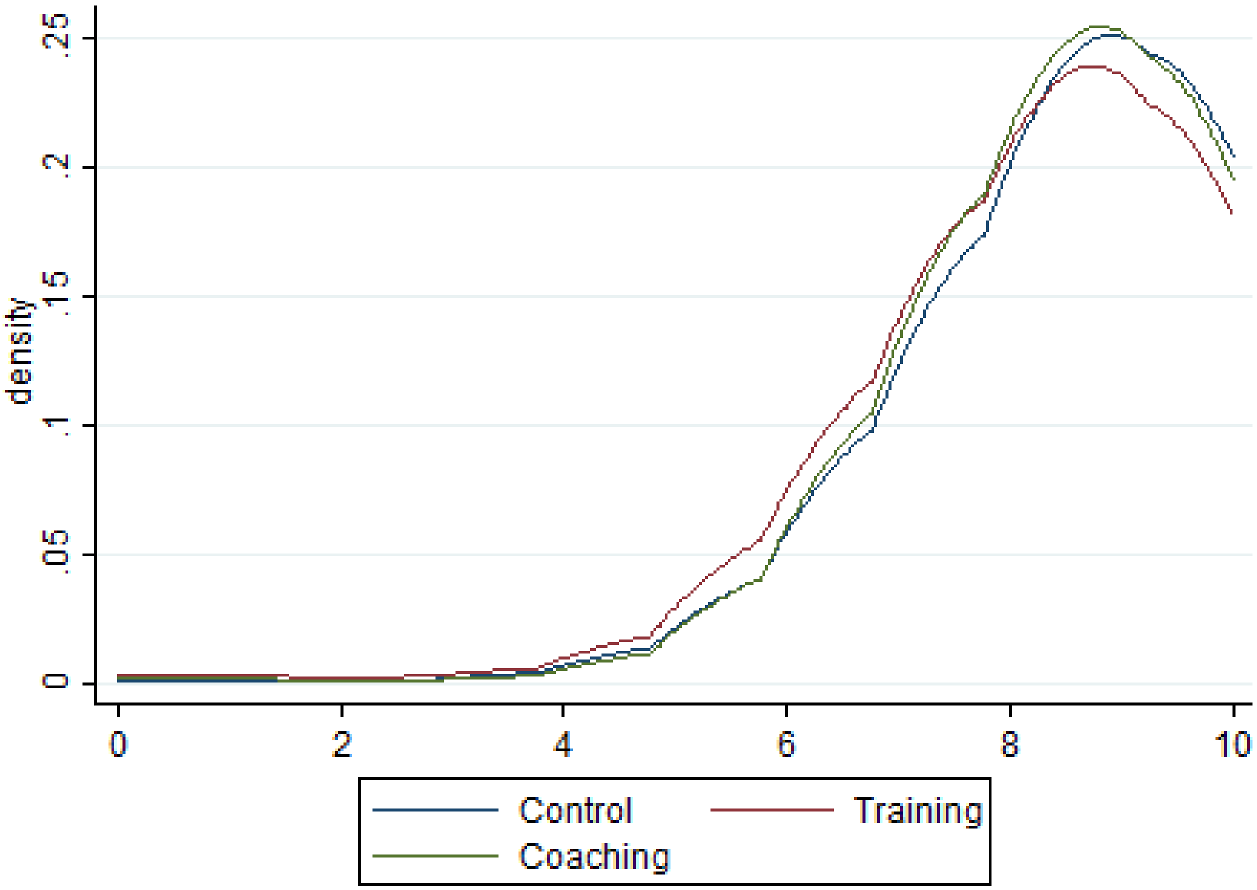


Figure A.3: Baseline distribution - letters correct

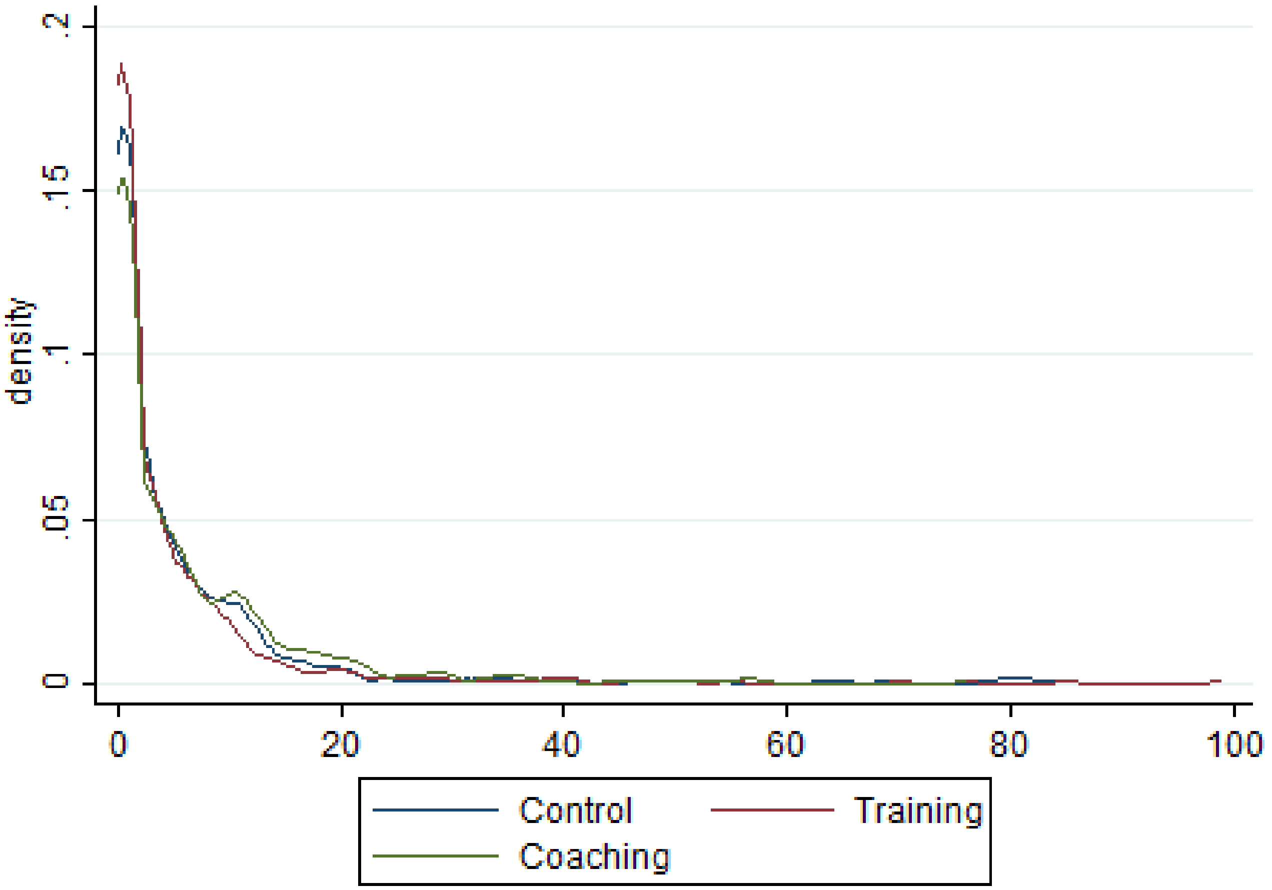


Figure A.4: Baseline distribution - digit span

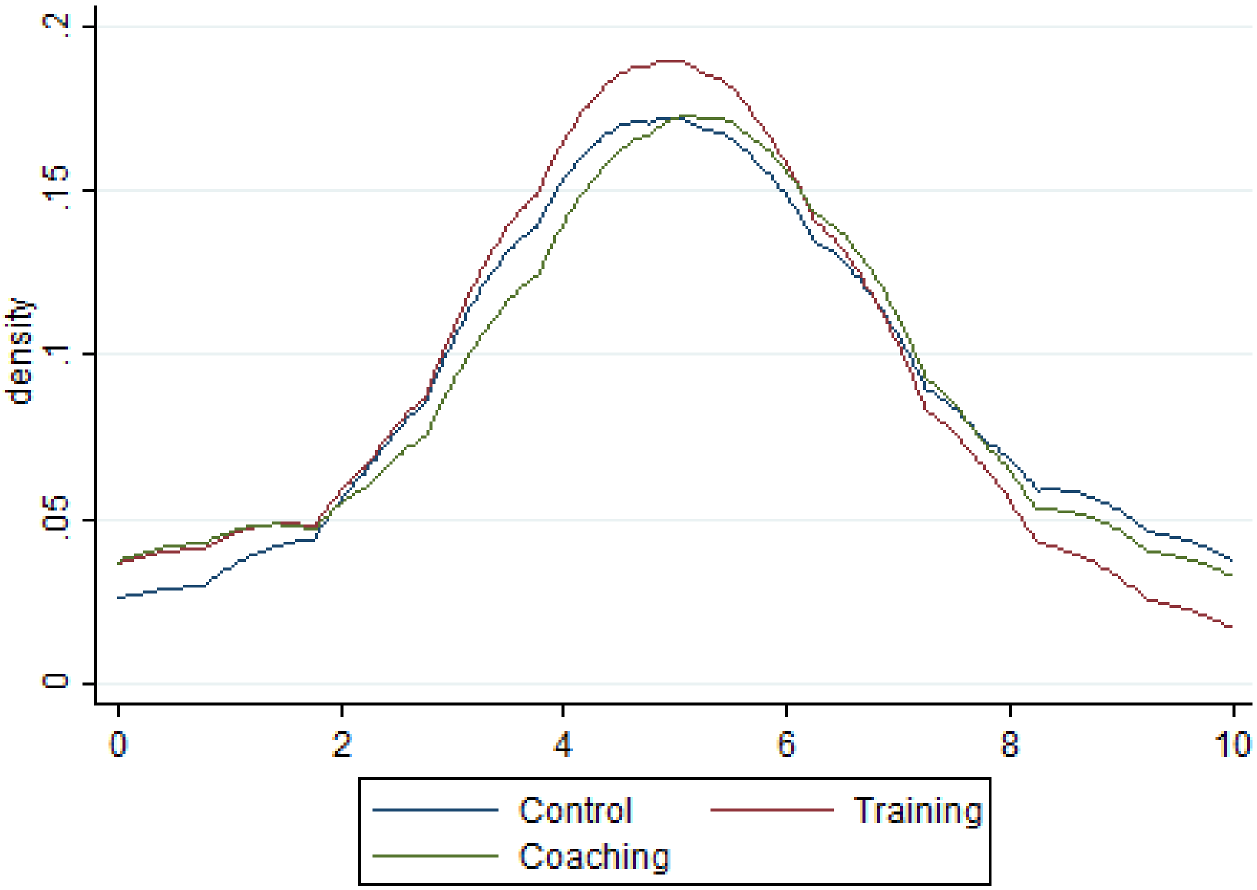


Figure A.5: Baseline distribution - phonemic awareness

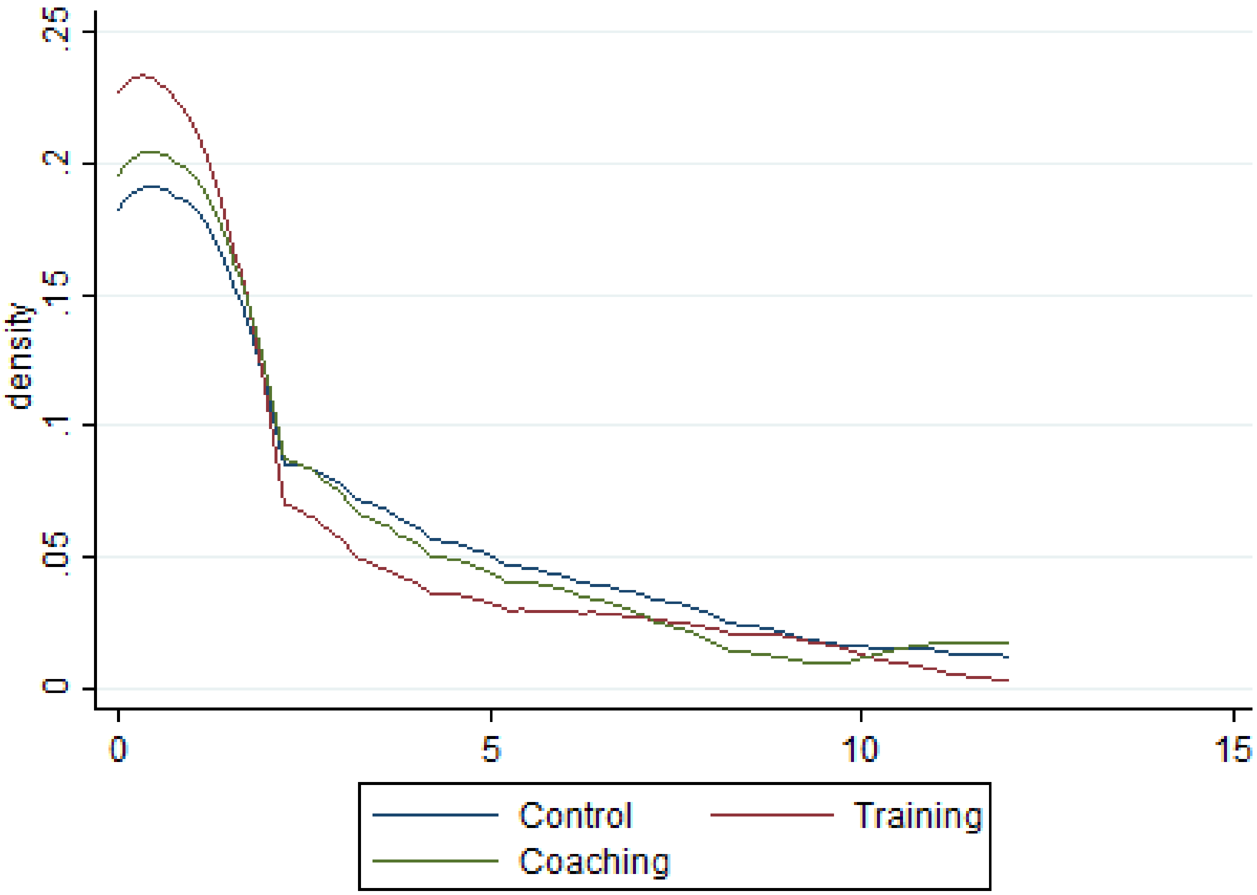


Figure A.6: Baseline distribution - words correct

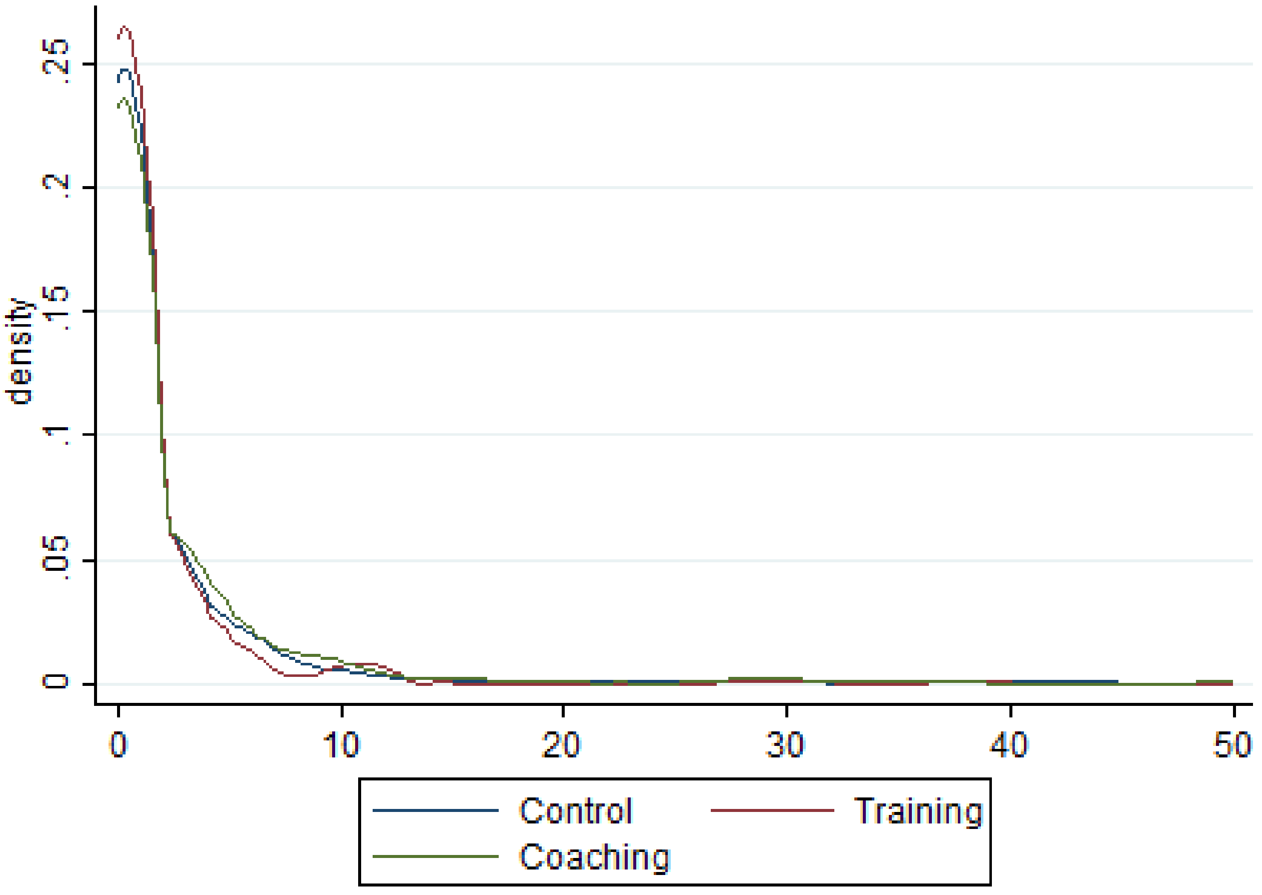


Figure A.7: Baseline distribution - comprehension test

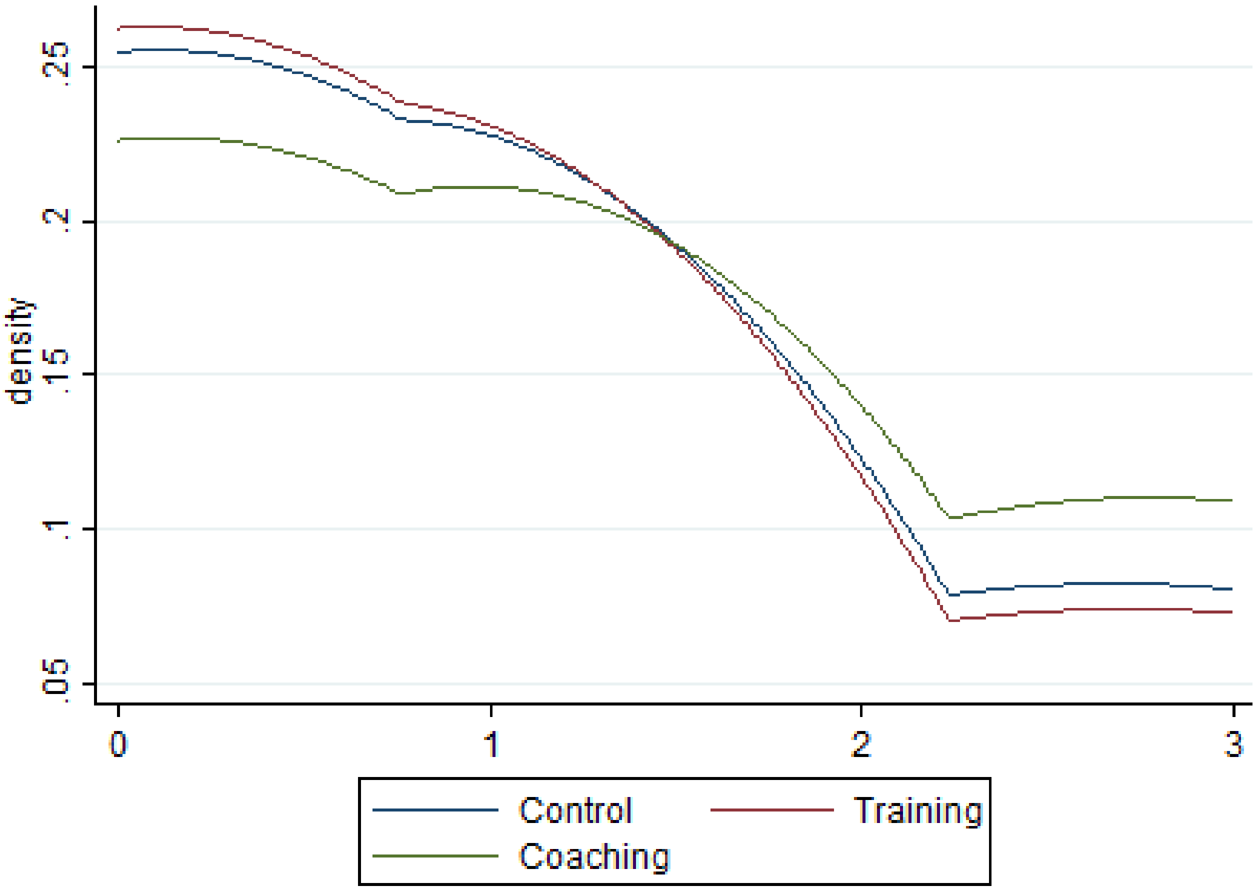


Figure A.8: Unconditional cumulative distribution of endline reading proficiency by treatment status

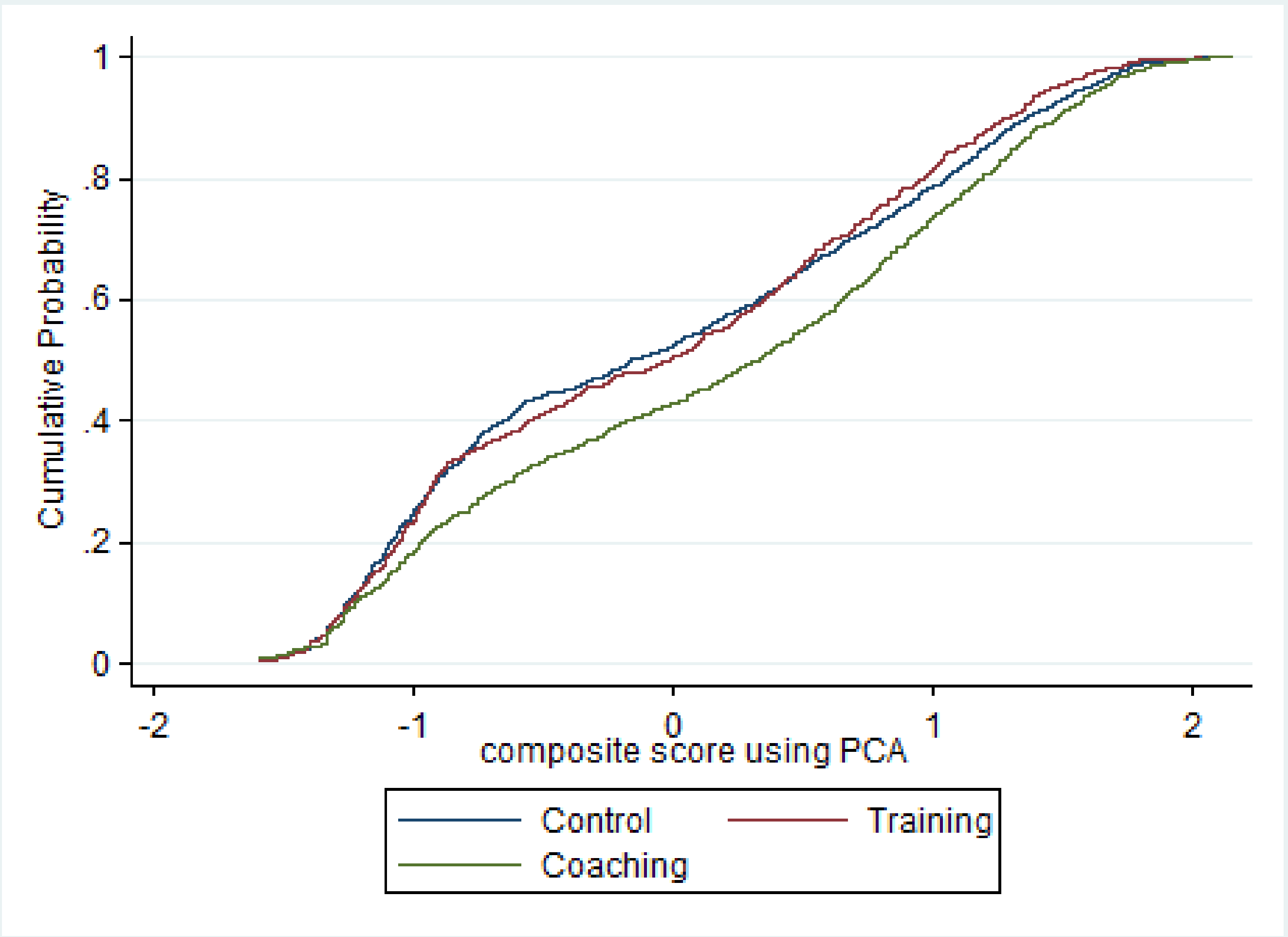
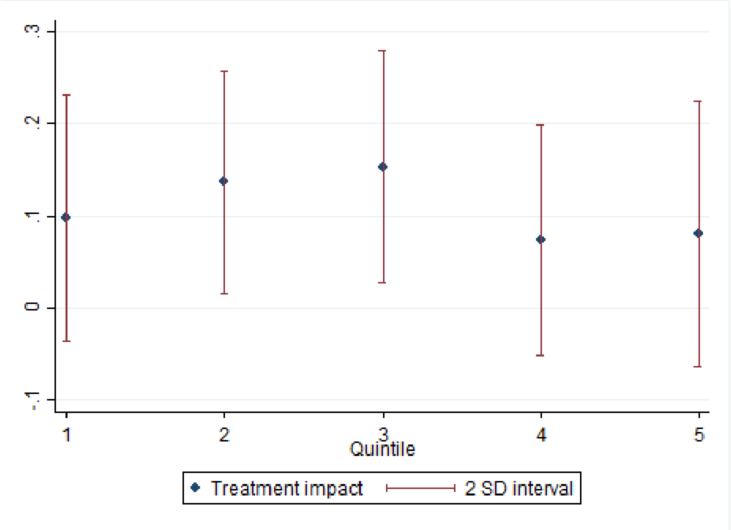
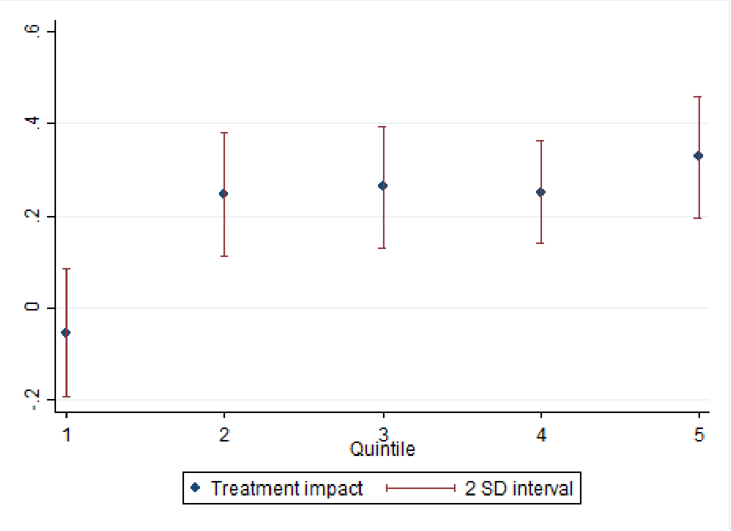


Figure A.9: Impacts by baseline reading proficiency



(a) Training



(b) Coaches