The Impact of Teacher Effectiveness

on Student Learning in Africa

Julie Buhl-Wiggers, Jason T. Kerwin, Jeffrey A. Smith and Rebecca Thornton [1]

April 2017

Abstract

Teaching quality is known to be critical for students' education and life prospects in developed countries. However, little is known about how teacher quality affects student learning in Africa. This paper presents the first estimates of teacher value-added from an African country, using data from a school-based RCT in northern Uganda. Exploiting the random assignment of students to classrooms within schools, we estimate a lower bound on teacher effects. A 1-SD increase in teacher quality leads to at least a 0.14 SD improvement in student performance on a reading test at the end of the year. Shifting teachers from the 10th to the 90th percentile of quality increases performance by 0.36 SDs –comparable to the most effective education interventions conducted in Africa. Our results also suggest that an increase in teacher quality can make other education interventions more efficient.

[1] Buhl-Wiggers: Department of Food and Resource Economics, University of Copenhagen (julie@ifro.ku.dk); Kerwin: Department of Applied Economics, University of Minnesota (jkerwin@umn.edu); Thornton: Department of Economics, University of Illinois (rebeccat@illinois.edu); Smith: Department of Economics, University of Michigan (econjeff@umich.edu).

1. Introduction

Teachers are important. Extensive evidence from developed countries shows that teacher quality has large effects on children's success in school and in adulthood, especially when exposed to quality teaching at young ages (Chetty et al. 2011, Chetty, Friedman, and Rockoff 2014). The evidence of the importance of teachers is consistent with research in developing countries, which finds that the interventions that are most effective at improving learning are those that focus on improving teacher training and reforming pedagogical approaches (Glewwe *et al.* 2014; Kremer *et al.*, 2013; McEwan 2014; Ganimian and Murnane 2014; Evans & Popova 2016). Yet, direct evidence on the effects of teaching quality in Africa is scant. Such evidence is much-needed: if variation in teaching quality drives large changes in student performance, there is scope for policymakers and administrators to improve learning by emulating the training of the most effective teachers or providing quality teacher support and mentoring. If teaching quality matters for student learning, but does not vary much across classrooms in Africa, policy should focus instead on recruiting more able teachers, and fundamentally changing how teachers are trained. Lastly, if we find teacher quality does not have large impacts on student learning, investing in other inputs first or simultaneously, may be necessary to improve educational outcomes.

This paper presents the first value-added estimates of teacher quality from an African country and among the first in a developing country. We utilize panel data from a randomized evaluation of a mother-tongue literacy program implemented in grades 1 to 4 in northern Uganda – the Northern Uganda Literacy Program (NULP). The program provided primary schools with intensive teacher training and support, scripted lesson plans, and revised learning materials. It began in a small number of pilot schools in 2010, where the materials and delivery of the program was tested and refined. A four year randomized evaluation of the program began in 2013; the first wave of the evaluation was conducted in 38 schools and in 2014 the evaluation was scaled up to cover 128 schools (Kerwin & Thornton 2015). Our analysis uses data from all four years of the evaluation exploiting the fact that students were randomly assigned to classrooms within schools in 2013 and 2016.

We estimate teacher effects using the value-added model approach, which takes student's prior achievement into account to control for variation in initial conditions and then estimates the increase in learning attributable to a specific teacher. The variation in these teacher effects is then interpreted as the variation in teacher quality.

To test whether the teacher value added is biased we follow Kane and Staiger (2008) and use the estimated teacher value added under non-random assignment to predict test scores under random assignment. In 2013 and 2016 we randomly assigned teachers to students in all grades. In the previous year the teachers were assigned in the usual manner. This means that we have adjacent cohorts where some was not necessarily randomly assigned to teachers whereas others were. We cannot reject that some sorting of students to teachers are biasing the results in years with non-random assignment. Therefore we test if average baseline scores are the same across streams within each school and only include schools where we cannot reject that teachers were as good as randomly assigned.

A large body of literature has estimated teacher effects in the United States and finds fairly consistent evidence that teachers are an important part of explaining the variation in test scores. This conclusion holds even when considering only variation in teacher quality within schools, and ignoring across-school variation. The estimated effect of a one-standard deviation increase in teacher effectiveness from schools in the United States, varies from 0.11 to 0.26 standard deviations of test scores (Hanushek and Rivkin 2010).

Little is known about how consistent the variation is between settings, as studies estimating teacher effects in developing countries are scarce. Among private secondary school teachers in India, Azam and Kingdon (2015) find that a one standard deviation improvement in teacher value-added increased test scores by 0.366 standard deviations. Talance (2015) uses panel data in grades 3 to 5 in Pakistan and finds that a one standard deviation increase in teacher quality increases student performance by 0.626 standard deviation. This is substantially higher than the results found in the U.S. In contrast, Araujo *et al.* (2016) find a one standard deviation increase in teacher value added increases test scores by 0.09 standard deviation among kindergarteners in Ecuador, which is quite similar to the results found in the U.S. No previous study has estimated value-added models of teacher ability in Africa; this paper is the first to do so.

Our lower-bound estimates of the teacher effects are that a one-standard deviation increase improves test scores by 0.14 standard deviations. These lower-bound estimates are derived from within-school variation, corrected for sampling variation. Shifting a teacher from the 10th to the 90th percentile causes a 0.36 standard deviation improvement in student performance. The data we use come from in northern Uganda, a very poor, rural setting with limited resources and teacher support, teaching materials are scarce, and class sizes are high - on average 80 pupils per teacher. Our results are important because they show that effective teachers are important for student success even in the resource-constrained conditions common in schools across rural Africa.

Our study allows us to assess the causal impact of improved support and training on teacher ability, using the fact that the NULP was randomized across schools. While the NULP intervention raises performance for all classrooms, it has an outsized impact for the most-effective teachers – and so increases the spread of classroom value-added. In the control group a one standard deviation increase in classroom value-added leads to an increase in performance of 0.13 SDs. For comparison, in the full program schools a one standard deviation increase in classroom value-added leads to an increase in performance of 0.22 standard deviations.

Given the substantial variation in quality, what are the characteristics of high-value add teachers? Using data from classroom observations and teacher surveys, we are able to describe what teacher characteristics and behavior are correlated with higher value-added measures. We find no correlation between teacher quality and teacher characteristics, but suggestive evidence that more effective teachers have more planned lessons, encourage participation and observe and record performance.

This paper has several implications; first teachers *do matter* in a low-resource context with several challenges in regard to quality education such as Uganda. Second, better teachers can make other interventions more effective implying an added benefit of shifting the worst teachers to the level of the most effective teachers. Third, observed teacher characteristics are not sufficient to measure teacher quality and thus more research is needed on who the most effective teachers are and how to recruit, train and support teachers.

2. Setting and Intervention Details

2.1 Primary Education in Uganda

Primary education in Uganda consists of seven years of education with a schooling starting age of six. To date the vast majority of Ugandan children have attended school at some point in time and the net enrollment rate is above 90% for both boys and girls. Despite this improvement in access issues regarding late enrolment, repetition and early drop out are still major challenges throughout the country, leading to many children being over-aged for grade. In order to graduate students must take the Primary School Leaving Exam (PSLE) and only about 54% reach this level (2013, World Bank WDI).

Since 1997, primary school has officially been free of charge, however, as resources are scare many schools still depend on contributions from parents, thus *defacto* school fees are common and students whose parents are not able to meet these contributions are often sent home. The reform of 1997 was successful in getting children into school. Yet, the large influx of children and limited resources has created raising concerns about diminishing school quality.

In 2007, the government of Uganda implemented a new curriculum in order to address these low education standards. This new curriculum, induced two main changes; Shifting the language of instruction from English to local language (11 different languages of instruction throughout the country) in lower primary (grades 1 to 3) and using a thematic curriculum instead of the traditional subject based. Despite this change Uganda still struggles with severe educational problems. 15 to 30 percent of all grade 7 students leave primary school without basic numeracy and literacy skills, respectively, where basic competencies mean reading a short story, with two comprehension questions for literacy and mastering division for numeracy (Uwezo, 2014). These findings are confirmed in a recent study which finds the vast majority (94% of children in government primary schools) could not read simple paragraph in English and infer meaning from it. Moreover, 54% could not order numbers correctly, 47% could not add double digit numbers and 76% could not subtract double digit numbers (Bold *et al.*, 2017). These numbers confirm that the Ugandan educational system has major learning challenges even 10 years after the latest attempt to reform the quality of the primary education system.

2.2 Teachers in Uganda

In order to become a qualified teacher in Uganda one must obtain a Grade III Teacher Certificate, which require two years of pre-service teacher training after four years of secondary school (O-level). For teachers already licensed and teaching in primary school the Grade III Teacher Certificate can be obtained through three years of in-service training. After obtaining the Grade III Teacher Certificate teachers can move on to obtain the Grade V Primary Certificate after two years of in-service training (MoES, 2013).

According to the Ministry of Education and Sports in Uganda 12.7 % of the primary school teachers was unqualified (not having a Grade III teacher certificate) to teach primary school in 2010. Among the qualified teachers, weakness in classroom pedagogy is still an issue as pre-service education is of poor quality with little transferability to the classroom (Hardman *et al.*, 2011).

Assessing the subject and pedagogical knowledge of teachers Bold *et al.* (2017) find that 16% of the teachers have minimum knowledge in language, 70% have minimum knowledge in math and only 4% have minimum pedagogical knowledge. In regard to classroom practices most teachers give positive feedback, but only half or less ask a mix of lower and higher order questions, plan the lessons or introduce and summarize the lesson. Very few teachers engage in all of the above practices (5%).

These weaknesses have led to a larger focus on in-service education and especially Continues Professional Development (CPD) which systematically update competences that teachers requires in the classroom. The CPD program is coordinated by the primary teachers' collages through Coordinating Center Tutors (CCTs). Tutors are typically recruited from experienced teachers and head teachers. They are responsible for providing workshops on Saturdays and during the school holidays and school-based support such as classroom observations and feedback to teachers and head teachers. However, one of the main challenges is to improve the technical capacities of the CCTs as much of the training they receive is too short to enable them to develop their own understanding of various teaching approaches and methods to best mentoring other teachers (Hardman *et al.*, 2011).

In addition to poor knowledge and pedagogical skills low levels of effective teaching time is a severe issue. Even though the average scheduled teaching time is around 7 hours a day effective teaching time is only 3 hours a day. This discrepancy is due to almost 60% of the teachers being absent from class leading to almost half of the classroom being orphaned (Bold *et al.*, 2017).

Teacher recruitment is administered at the central level based on the amount of funds available for teacher salaries. Vacancies are identified at the school level by the head teacher. These vacancies are then sent to the District Education Officer who compiles all the vacancies in the district which is then sent to the central government. As teachers are scarce, the first step is to re-allocate teachers from schools with a surplus of teacher to schools with a lack of teachers within the same district. When this is done the total amount of teachers that needs to be recruited is calculated from the available funds. As the government budget does not allow for an adequate number of teachers some schools are obliged to recruit teachers off payroll and pay them using resources mobilized by the school (usually from parents through mandatory school contributions). It is estimated that 2% of the teachers are off pay-roll (MoES, 2013).

Teacher attrition is estimated to around 4% and the two major causes are resigned (21%) and dismissed (14%) suggesting that the working environment is characterized by dissatisfaction of the teachers and issues related to ethics and teacher behavior. A survey conducted by the Ministry of Education and Sports does indeed show low levels of job satisfaction among primary teachers and the vast majority would like to leave the teaching profession within two years (MoES, 2013). The main cause of job dissatisfaction stated is low salary, which is minimum 511,000 Ugandan shillings per month (corresponding to $150).

2.3 Northern Uganda Literacy Project (NULP)

The program we study, the Northern Uganda Literacy Project (NULP), is a literacy intervention developed in response to the educational challenges facing northern Uganda. The NULP was designed by a locally owned educational tools company called Mango Tree Educational Enterprises Uganda (henceforth Mango Tree), and aims to increase early childhood literacy skills through a mother-tongue-first instructional approach and extensive teacher training and support. The project is based in the Lango sub-Region, where the vast majority of the population speaks Leblango. The NULP model involves a revised curriculum for grades 1 to 3 that focuses on mother-tongue-first instruction and moves at a slower pace to ensure the acquisition of fundamental literacy skills. This curriculum is paired with detailed, scripted teacher guides that lay out lesson plans for teachers and intensive teacher training and support, as well as primers and readers for every student, and slates and chalk for students in grade 1. A scripted approach like the NULP's has been used with some success in the United States, but has proven controversial among American teachers (Kim and Axelrod, 2005). It is particularly well-suited to teaching literacy in the Lango sub-Region, an area where teachers are often inadequately trained. The NULP's fixed, scripted lessons also fit into a fixed weekly schedule. This helps keep both teachers and students on track, giving them an easy-to-remember and easy-to-use routine for literacy classes. Among the program schools the teachers receiving the program was depended on which grade they were teaching. In 2013 and 2014 all P1 teachers received the program and in 2015 and 2016 all P2 and P3 teachers received the program, respectively.

3. Sample, and Data

3.1 Sample

Our dataset consists of four cohorts of children, followed from P1 to either P2, P3 or P4 depending on the year they started P1 - 2015, 2014 or 2013, respectively. There are two main samples we work with in our analysis. The first sample includes all teachers available and is used to estimate classroom effects. The second restricts the sample to teachers who are teaching in multiple years as this is needed in order to estimate teacher effects. We describe these samples of schools, teachers, and students below. Table 1 presents the sample statistics for each of the two samples.

[Table 1 about here]

(i) Schools

Schools were sampled for the study in two phases: an initial RCT in 2013, and a larger RCT in 2014 which carried on in 2015 and 2016. In 2013, 38 eligible schools were selected to be part of the RCT. To be eligible schools had to meet a set of criteria established by Mango Tree, the most important being that each school needed to have exactly two P1 classrooms and teachers[2]. In 2014 the program was expanded to 90 additional schools for a total of 128 schools. The eligibility criteria for these new 90 schools were slightly different, and less stringent.[3]

(ii) Teachers

Our sample of teachers is largely grade-specific rather than cohort-specific. Since Ugandan schools practice social promotion, students in principle always advance to a new grade each year. Therefore, our sample of classrooms includes the same cohort of teachers each year, but a new cohort of students. In the initial 38 schools (and hence all of the 2013 data) we have two teachers in every school except one. Moreover, when restricting the students per classroom to be minimum 10 students we lose two teachers, leaving us with a total of 73 teachers.

In 2014, we have 122 new P1 teachers from the 90 new schools and 22 new P1 teachers from the original 38 schools that entered the sample. In addition, in the original 38 schools, 44% of the

[2] One school did not have two grade 1 classes and is removed from the analysis using school fixed effects. Other eligibility criteria include including: being located in one of five specific school districts (coordinating centres), having, having desks and lockable cabinets for each P1 class, a student-to-teacher ratio in P1 to P3 of no more than 135 during the 2012 school year, located less than 20 km from the headquarters of the coordinating centre, accessible by road year round, had a head teacher regarded as "engaged" by the coordinating centre tutor, and not having previously received support from Mango Tree.

[3] Criteria in 2104 include: having desks and blackboards grade P1 to P3 classrooms and having a student-to-teacher ratio of no more than 150 students during the 2013 school year in grades P1 to P3.

teachers are reassigned to new grade or different schools due to turnover, promotion, and so forth. When restricting the class size to minimum 10 students we lose 7 teachers, leaving us with a total of 178 P1 teachers.

In 2015, 55 % of the P1 teachers in 2014 were still teaching P1, 10 % were teaching P2 or P3 and the remaining 35 % were teaching higher grades or not found. In addition, two teachers from the 2013 sample re-entered and 16 new teachers entered the sample, leaving us with 148 P1 teachers. For P2 we have 171 teachers and for P3 we have 46 (P3 is only tested in the original 38 schools).

In 2016, 61 % of the P1 teachers in 2015 were still teaching P1, 3 % were teaching P2 or P3 and the remaining 37 % were teaching higher grades or not found. In addition 31 teachers from 2013 and 2014 re-entered and 26 new teachers entered the sample, leaving us with 151 P1 teachers. For P2 40% of the P2 teachers in 2015 still taught P2, and for P3 37 % still taught P3.

All in all, we have 712 teachers across all years and grades of these 279 (or 39%) are teaching in at least two years.

(iii) Students

In 2013, 50 P1 students were sampled at random from each of the 38 schools based on enrollment lists collected at the beginning of the school year. The sample was stratified by classroom and gender, resulting in 25 students per classroom. In 2014, 2015 and 2016 this initial sample of P1 students was retained, and tracked into P2[4], P3 and P4, respectively. By the end of P4 the attrition rate was close to 60%.

In 2014, we added a new cohort of P1 students to the study. Among this new cohort, 100 P1 students were sampled at random from each of the 128 schools[5]. As the first cohort this cohort was also tracked into P2 and P3 in 2015 and 2016, respectively. The attrition rate in P3 was close to 50%

[4] In this version of the paper we don’t have teacher information from P2 in 2014 and thus is not included.

[5] The sampling procedure differed slightly across the original 38 schools and the 90 added in 2014 due to logistical constraints. In the 38 schools that had participated in 2013, an initial sample of 40 P1 pupils was drawn at baseline 2014, and then 60 students were added at endline following the same procedure as was used to add pupils to the P2 sample. In the 90 new schools, the initial sample was 80 pupils and 20 top-up pupils were added at endline. This difference in the numbers of students sampled at the beginning of the was due to the organizational difficulty of handling large numbers of students in the original 38 schools, since they also had a sample of 50 P2 students. For the end of year testing in 2014, this difficulty was addressed by hiring additional enumerators.

In 2015, a third and smaller cohort, 30 randomly sampled P1 students in each school, was added and tracked into P2 in 2016. The attrition rate in P2 was 32%. In 2016, the fourth cohort was added, by randomly sampling 60 P1 students in each school.

3.2 Randomization

(i) Assignment of NULP to schools

To assess the impact of the NULP on student learning, we conducted a multi-year, randomized evaluation of the program (described in more detail in Kerwin & Thornton 2015). Of the 38 schools in 2013 and 128 schools in 2014, the evaluation assigned each to one of three study arms: 1) Full-cost, 2) Reduced-cost, and 3) Control. In the Full-cost group, schools received the original NULP as designed by and delivered by Mango Tree and its staff. In the Reduced-cost group, some of the materials (slates and chalk) were eliminated, training was conducted through a cascade model led by government employees (Ministry of Education staff) rather than Mango Tree staff, and teacher received fewer support visits, again from government employees. Schools in the Control group did not receive the literacy program.

To randomize, schools were grouped into stratification cells of three schools each. Each stratification cell had its three schools randomly assigned to the three different study arms via a public lottery.

(ii) Assignment of students to classrooms and teachers

Our research design takes advantage of the fact that students were randomly assigned to teachers. For the 2013 and 2016 classes students were randomly assigned to classrooms by providing the head teacher in each school with blank rosters that contained randomly-ordered classroom assignments. Each head teacher then copied the names of all students from his or her own internal student list onto the randomized roster in order, which generated a randomized classroom assignment for each student. Students that enrolled late were added to the roster in the order they enrolled, and thus were randomly assigned to classrooms as well. Compliance with this procedure was verified by having field staff compare the original student lists to the randomized rosters, and also by asking the head teachers what they did. In order to test compliance we take the approach suggested by Horváth (2015) and test differences in baseline score means between classrooms within schools and grade level for each year. We found that a few schools had classrooms with baseline difference between classes and we excluded those when assessing the degree of bias present[6].

In 2014 and 2015, head teachers were not given explicit instructions on how to assign students or teachers. In general, the way assignments are made is specific to each school, and depends on the approach used by the school's head teacher. In order to assess the degree of sorting present in these years we also test the differences in baseline score means between classrooms within schools and grade level for each year. We find that for most schools we cannot reject that children were as good as randomly assigned to teachers.

3.3 Data

Our data consists of 20,190 children and 30,370 children-by-year observations. Summary statistics on test scores and basic demographics is presented in Table 2. We have restricted our sample to only including classes with more than 10 students per teacher and due to this restriction we lose four schools in 2014 and one school in 2015. The average age at the end of P1 is around 7.5 years, which means that some students start later than the official schooling starting age of 6; 50 percent of the students are girls.

[Table 2 about here]

(i) Learning Outcomes

Our outcome of interest come from the Early Grade Reading Assessment (EGRA) which is an internationally recognized exam to assess early literacy skills such as recognizing letters, reading simple words and understanding sentences and paragraphs (Dubeck & Gove., 2015; Gove & Wetterberg, 2011; Piper, 2010; RTI International, 2009). We use a validated adaptation of the EGRA to Leblango, which covers six components of literacy skills: letter name knowledge (LN), initial sound identification (IS), familiar word recognition (FW), invented word recognition (IW), oral reading fluency (ORF), and reading comprehension (RC). In order to measure overall

[6] See Appendix E for distributions of the P-values

performance we construct a principal components score index in the following way. First, we normalize each of the test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006). This procedure is done separately for each year and grade.[7] [8]

Tests are administered at the beginning and end of year in both 2013 and 2014. In 2013, of the 1,755 students for which we have both classroom information and beginning-of-year test scores, 1,357 students were present for the endline exams (77%). In 2014, of the 5,201 students with classroom information and beginning of year test scores, 4,409 were present for the endline (85%). In 2015 and 2016 the tests were only administered at the end of the year. As the vast majority of P1 students (90%) score zero when tested at baseline in 2014 we find it reasonable to set the baseline score for P1 in 2015 and 2016 to zero[9]. This means that for P1 students the value-added is from no skill to the skills obtained at the end of the year. In 2015, we have 3,423 P1 students, 5,571 P2 students and 1,210 P3 students. This corresponds to 61% (P2) and 55% (P3) of the students tested at the end of 2014 in P1 and P2, respectively. In 2016, we have 6,795 P1 students, 2,241 P2 students, 4,533 P3 students and 851 P4 students. This corresponds to 63% (P2), 55% (P3) and 64% (P4) of the students tested at the end of 2015 in P1, P2 and P3, respectively. Attritors are more likely to be younger, performing worse on the baseline test and not being enrolled in a full treatment school.

(ii) Teacher Characteristics and Teaching Practices

Data on teacher characteristics are obtained from a teacher survey conducted in the beginning of 2013 and 2014. From this survey we have information on both individual and household characteristics. We also conducted a three-question Raven's Standard Progressive Matrices (SPM) test to measure fluid intelligence, as well as asking a range of questions in social science, science, math and language. Table 2 shows that the average teacher is around 41 years old, has 15 years of education (which corresponds to two years of post-secondary education) and has a total score of 2.27 out of 3 on the SPM test, or 75% correct. This would put the average teacher at around the 50th

[7] See Appendix A for the results of the principal components analysis. See Appendix B for the distributions of the endline PCA scores by grade level.

[8] Some students, 31 in 2013 and 993 in 2014, are missing at least one component of the beginning-of-year test score, which results in a missing beginning-of-year test score when we construct the PCA index. Our results are robust to alternative methods of index construction, where we only lose the test score if all components are missing

[9] See Appendix C for the distributions of the baseline subtest in 2013 and 2014.

percentile of the US adult distribution on the full 60-item SPM (Bilker et al. 2012). Roughly one third is women and the vast majority is married.

The new teachers included in 2014 are slightly less educated (14 years of education), score lower (1.87) on the SPM test and are more likely to be women. These differences are likely to be driven by the fact that the expansion in 2014 induced less stringent eligibility criteria, thus including more rural and disadvantaged schools.

In 2013 we also conducted in-person observations of each classroom in the study. These classroom observations were done by experienced enumerators and measured teacher and student demeanor, discipline, interactions between teachers and students, the use of Leblango and English, and time spent on teaching. The goal of the classroom observations was to measure teacher behaviors that are relevant to teaching literacy and might be predictive of the successful implementation of the NULP instructional model. We therefore did not use a standard rubrics such as the CLASS, but instead designed our own tool to capture the behaviors of interest. The observations were conducted three times that year, in July, August and October. Each 30-minute lesson was broken up into three 10-minute observation windows; for each block the enumerator ticked of boxes to indicate the actions which occurred during that time period. Even though these observations were done in a small sample of teachers (61 teachers) it gives an indication of what is going on in the classrooms of our sampled children.

The vast majority of teachers refers to the Teachers Guide, moves freely around the classroom, calls on the individual and encourages participation. However, most of them do not record performance of the students. If we look at how classroom practices differ between different types of teachers we see that there is no significant difference in classroom practices between teachers of different age or gender. When we look at years of education and ability (measured as SPM) we see that teachers in the top of the distributions (above the 90th percentile) are doing significantly better on some classroom practices. High educated teachers are more likely to call on the individual and less likely to not participate. Moreover, teachers at the top of the ability distribution are more likely to bring pupils back on task, record performance and less likely to have a negative demeanor.

4. Conceptual Framework and Empirical Strategy

We turn next to our conceptual framework and empirical strategy for estimating the effects of teachers on student learning.

4.1 Conceptual Framework

Learning is a complex, cumulative process that depends on students' cognitive and non-cognitive ability as well as their current and prior home environment, teacher quality, peers and other school-specific factors amongst others. Wolpin and Todd (2003) describe the canonical model of the production of the learning process as follows:

$$(1) \qquad Y_{icst} = Y_t[\boldsymbol{X}_{ics}(t), \boldsymbol{S}_{ics}(t), \boldsymbol{C}_{\boldsymbol{ics}}(t), \theta_{i0}, \varepsilon_{icst}]$$

where Y_{icst} is a measure of achievement for child *i* in classroom *c* in school *s* at time *t*. Acquisition of this knowledge is then modelled as a combination of cumulative family-supplied inputs ($\boldsymbol{X}_{ics}(t)$), cumulative school-supplied inputs ($\boldsymbol{S}_{ics}(t)$) such as school management etc., cumulative classroom inputs such as the teacher ($\boldsymbol{C}_{ics}(t)$) and genetic endowments (θ_{i0}). ε_{icst} allows for measurement error in the achievement measure. Y_t allows the impact of all factors to depend on time and thus the age of the child. As data on this entire process is rarely, if ever, available, many scholars have sought alternative ways of estimating the determinants of learning. One approach in economics is the value added model, which takes prior student achievement into account to control for variation in initial conditions. Treating the arguments in (1) as additive separable and assuming that the parameters are not varying with age, equation (1) reduces to:

$$(2) \qquad Y_{icst} = \beta_0 + \beta_1 Y_{icst-1} + \beta_2 \boldsymbol{X}_{\boldsymbol{icst}} + \rho_s + \lambda_{cs} + \varepsilon_{icst}$$

where, Y_{icst-1}captures previous family, school and individual factors as well as genetic endowments. ρ_s is the effect of the school such as skills of the principal etc. λ_{cs} is the effect of being in a specific classroom and thus λ_{cs} is an estimate of the increase in learning attributable to a specific classroom and teacher. The variation in these classroom effects is then interpreted as the variation in teacher quality.

4.2 Empirical Strategy

4.2.1 Classroom Effects

We start our analysis by estimating classroom effects using the "lagged-score" value added model presented in equation (2). In order to estimate λ_{cgst}, two issues arise: First, in our data teachers are perfectly nested within schools as we do not have any teachers that switch schools in our sample. Therefore we cannot separate out the classroom effect from the school effect by including school fixed effects. Second, the error term ε_{it} is likely to include individual heterogeneity such as learning speed. While $Y_{icgst-1}$ captures innate ability, more able children may also learn faster and if not controlled for will be captured by the error term. In effect, we estimate the following partially controlled equation:

$$(3) \qquad Y_{icgst} = \hat{\beta}_0 + \hat{\beta}_1 Y_{icgst-1} + \hat{\beta}_2 \boldsymbol{X_{icgst}} + \hat{\lambda}_{cgst} + \zeta_g + \tau_t + \hat{\upsilon}_{icst}$$

where

$$(4) \qquad \hat{\upsilon}_{icgst} = \rho_s + \theta_i + \varepsilon_{ics}$$

$\hat{\lambda}_{cgst}$ is estimated as a classroom fixed effect. To allow for interdependence within classrooms, we cluster the standard errors at the classroom level.

There are three factors potentially threatening the identification of our estimated classroom effects $\hat{\lambda}_{cgst}$ in equation (3). First, there may be school effects that covary with true classroom effects, $\hat{\lambda}_{cgst}$, such as school management, resources or other things that can influence school choice. Second, there may be individual effects that covary with true classroom effects, such as sorting of students to teachers based on parental influence or other unobserved characteristics. Third, sampling error. The estimated classroom effects are the sum of the true classroom effects and the estimation error that arises from the fact that we have small samples of students, thus:

$$(5) \qquad \hat{\lambda}_{cgst} = \hat{\lambda}_{cgst} + \frac{1}{N_{cs}} \sum_{i=1}^{N_{cs}} \varepsilon_{ics} \, ,$$

where N_{cs} is the number of students per class. As the sample gets small (fewer students tested per class) the sampling error gets large. This sampling error could overwhelm the signal, causing a few very low or very high performing students to strongly influence the estimated classroom effects, $\hat{\lambda}_{cgst}$.

We address each of these three potential threats to identification in turn in the following sections.

(i) Purging the school effects

When estimating equation (3) we use both within- and between-school variation which means that the estimated $\hat{\lambda}_{cgst}$ picks up both classroom effects and school effects that covary with the classroom effects, $\hat{\lambda}_{cgst} = (\lambda_{cgst} + \kappa\rho_s)$. To overcome this problem we rescale the classroom effects $\hat{\lambda}_{cgst}$ to be relative to the school mean and thereby only consider the within-school variation in the classroom effects (Slater, 2012; Araujo, 2016). This means that the rescaled classroom effects become:

$$(6) \qquad \hat{\gamma}_{cgst} = \hat{\lambda}_{cgst} - \frac{\sum_{c=1}^{C_s} N_{cs}\hat{\lambda}_{cgst}}{\sum_{c=1}^{C_s} N_{cs}}$$

where $\hat{\gamma}_{cgst}$is the demeaned classroom effect, C_s is the total number of classrooms within the school *s* and N_{cs} is the total number of students in classroom *c* and school *s*. This approach nets out all school level factors and thereby provide a lower bound to the degree of variation.

(ii) Sorting of students to teachers

Our empirical strategy to estimate classroom effects relies, in part, on the fact that students were randomly assigned to teachers in 2013 and 2016. One threat to the validity of this approach would be if students systematically switched classrooms during the year, or if student attrition was correlated with teacher ability. In 2014 and 2015 students were not randomly assigned to teachers, so we cannot rule out that the classroom effects are biased due to sorting of students to teachers. Including prior test scores as controls reduces this concern, but does not eliminate it: we would not expect prior test scores to be a fully comprehensive measure of ability and other unobserved characteristics.

To assess the degree of bias due to non-random assignment we follow Kane and Staiger (2008), Chetty *et al.* (2014a) and others, and test whether the estimated classroom effects from non-random can accurately predict the mean test scores of the students who are randomly assigned to the classrooms. In practice, we use the sample of pupils with randomly assigned teachers and purge the endline test scores of observed characteristics and obtain the residuals:

$$(7) \qquad Y_{ics} = \beta_0 + \beta_1 Y_{ic-1s} + \beta_2 \boldsymbol{X_{ics}} + \rho_s + \epsilon_{ics}$$

where Y_{ics} is the end-of-year test score. Y_{ic-1s} is the beginning-of-year test score, $\boldsymbol{X_{ics}}$ is a vector of student characteristics including age and gender, ρ_s is a vector of school indicators and ϵ_{ics} is the error term. Then we regress these residuals, ϵ_{ics} on the estimated demeaned classroom effects from non-random assignment.

$$(8) \qquad \epsilon_{ics} = \alpha\hat{\gamma}_{cgst} + \vartheta_{ics}$$

If we are unable to reject the hypothesis that α equals one, it suggests that the value-added measure is unbiased: the difference in test scores under random assignment equals the difference in the value-added measure.

(iii) Sampling variance

As described above, the estimated variance of the classroom effects is the sum of the true variance and the sampling variance. This is particularly problematic when we have a small number of student test scores in each class. To address this problem we analytically adjust the variance of the classroom effects following the approach suggested by Araujo *et al*. (2016).[10] For the within-school classroom effects we estimate the variance of the measurement error as $\frac{1}{C}\sum_{c=1}^{C}\left\{\frac{\left[\left(\sum_{c=1}^{C_s} N_{cs}\right)-N_{cs}\right]}{N_{cs}\left(\sum_{c=1}^{C_s} N_{cs}\right)}\sigma^2\right\}$, where σ^2 is the variance of the residuals, ε_{ics} from equation 3. N_{cs} is the number of students in classroom, c in school, s, C_s is the number of classrooms in school, s and C is the overall number of classrooms. Then we subtract that from the estimated variance of the demeaned classroom effects:

$$(9) \qquad \hat{V}_{corrected}\left(\hat{\gamma}_{cgst}\right) = V\left(\hat{\gamma}_{cgst}\right) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{\left[\left(\sum_{c=1}^{C_s} N_{cs}\right)-N_{cs}\right]}{N_{cs}\left(\sum_{c=1}^{C_s} N_{cs}\right)}\sigma^2\right\}$$

For the classroom estimates that also use between-school variation this expression reduces to:

$$(10) \qquad \hat{V}_{corrected}\left(\hat{\gamma}_{cgst}\right) = V\left(\hat{\gamma}_{cgst}\right) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{1}{N_{cs}}\sigma^2\right\}$$

4.2.2 Teacher effects

In principle the estimated classroom effects from equation (3) contain both a permanent teacher component as well as a transitory classroom component that captures; disturbances during testing, peer dynamics etc. When we have more than one year of data for the same teacher, under some circumstances it is possible to separate the teacher effect from the classroom effect. In order to

[10] The procedure is analogous to the Empirical Bayes approach. The difference is that the procedure proposed by Araujo *et al.* (2016) explicitly accounts for the fact that the classroom effects are demeaned within each school and that the within school mean may also be estimated with error. See the online appendix D of Araujo *et al.* (2016) for details.

obtain the teacher effects we use the demeaned classroom effects and estimate the following equation:

$$(11) \qquad \hat{\gamma}_{cgst} = \hat{\alpha}_0 + \hat{\hat{\delta}}_{cgs} + \omega_{cgst}$$

Where, $\hat{\hat{\delta}}_{cgs}$ is a vector of teacher indicators and can be interpreted as the permanent teacher component. These are our coefficients of interest when discussing the teacher effects. We test the degree of bias and correct for sample variation in the same manner as described above for the classroom effects.

4.2.3 Correlation with Teacher Characteristics and Behaviors

In order to describe the characteristics and behaviors of the most effective teachers, we show how these are correlated with our estimated value-added measures. For the teacher characteristics we estimate the following equation:

$$(12) \qquad \hat{\hat{\delta}}_{cgs} = \alpha_0 + \alpha_1 C_{cgs} + \upsilon_{cgs}$$

Where $\hat{\hat{\delta}}_{cgs}$ are our estimated teachers effects from equation (11), C_{cgs} is a vector of teacher characteristics and includes; gender, age, salary, years of schooling, number of correct answers on the SPM and if the teacher lives in the same village as the school. In addition, we provide graphical evidence of the correlation between various classroom behaviors and the estimated teacher effects ($\hat{\hat{\delta}}_{cgs}$).

5. The Impact of Teacher Effectiveness on Student Learning

5.1 Baseline results

Table 3 presents our baseline results estimated from equation (3) and (11) and shows several estimates of the classroom and teacher value-added measures summarized in terms of standard deviations of student performance on the endline exams.

[Table 3 about here]

The first column shows the results of estimating the classroom effects with all teachers available in our sample and the second column shows the results from reducing the sample to teachers with at least two years of data. The final column shows the results from estimating equation (11) and thereby obtaining the teacher level average of the classroom effects across years which can be interpreted as teacher effects.

Panel A shows the results from the naïve model which uses both between and with school variation to estimate the classroom and teacher effects. We find a substantial amount of variation between teachers. A 1 SD increase in teacher quality increases student performance with 0.43 to 0.52 SDs, when correcting for sampling error. Yet, as these estimates also include between school variation some proportion of the effect is likely to be due to sorting of teachers to schools. To identify the part due to the teacher we in Panel B limit the variation to only with-school and effectively only compare teachers between either stream or grade in the same year and school. Using this specification we still find substantial variation between teachers. The most restrictive results show that a 1 SD increase in teacher quality increases student performance with 0.19 SDs.

5.2 Bias and Results from Random Assignment

In Table 4, we assess the potential bias stemming from non-random assignment of students to teachers. The table shows the results separately from purging the school effect (column 2) or not (column 1). We see that the estimate in both columns are significantly different from one, suggesting that we cannot rule out that sorting of students to teachers are biasing our estimates.

[Table 4 about here]

To investigate if this potential bias is material we restrict our sample to include only classes where the students were randomly assigned to teachers and present the results in table 5. In column one we present the standard deviation of the classroom effects estimated from teachers that taught in either 2013 or 2016 in schools where we cannot reject that the average baseline score between two streams are the same. We find that a 1 SD increase in classroom effectiveness increases student performance by 0.16 SDs. As mentioned previously estimating teacher effects require at least two years of data from the same teacher. For the teachers teaching in 2013 we only have 12 teachers also teaching in

2016 which makes the sampling error overwhelm the estimated variance. One way to overcome this problem is include teachers teaching in 2014 and 2015 in schools where we cannot reject that the average baseline score between two streams are the same. Thus, we re-estimate the classroom effects including these teachers. As seen from column two the results are very similar to the results in column one supporting the claim that it is reasonable to include these teachers in order to estimate teacher effects. In column three we present the standard deviation of the teacher effects using teachers from all years in schools where we cannot reject random assignment. Overall, we see that both classroom and teacher effects are very similar across samples and that a 1 SD increase in teacher effectiveness increases student performance by 0.14 to 0.16 SDs.

[Table 5 about here]

Comparing the results in table 5 with the baseline results in table 4 we see that under random assignment (table 5) the standard deviation of the classroom effects is around 50% smaller implying that the direction of the bias is that higher quality teachers are matched with better students. As described in section 4 the teacher effects are estimated as the teacher level average of the classroom effects across years. If sorting is not systematically occurring year after year the teacher effects would be less prone to bias as the bias would be purged as a transitory year effect. Indeed the difference is smaller when comparing the standard deviation of the teacher effects between table 3 and 5. This suggests that a substantial part of the sorting is not systematically occurring year after year making teacher effects a reasonable measure of teacher effectiveness even in the absence of random assignment.

The corrected standard deviation of the within school teacher effects (table 5, column 3, panel B) is our preferred estimate of teacher effectiveness and can be interpreted as a lower bound as we only use with school variation. This means that a 1 SD increase in teacher effectiveness increases student performance by 0.14 SDs which implies that moving from a 10th-percentile teacher to a 90th-percentile one would mean a gain in student learning of 0.36 SDs.[11]

[11] These calculations assume the teacher effects are normally distributed, and hence use the standard normal distribution to compute that a move from the 10th to the 90th percentile of the distribution is a 2.564-SD change in teacher ability.

5.3 Robustness

In this section we address the robustness of our estimates. In particular we estimate if our estimates are robust to excluding all P1 students in 2015 and 2016. As mentioned in section 3.3 baseline scores were not collected in 2015 and 2016 which led us to impute all P1 baseline scores in 2015 and 2016 with the median P1 score in 2013 and 2014 (in principle zeros). While imputing the baseline scores for P1 in 2015 and 2016 allows us to retain a larger sample of teachers over time it also by implication introduce more measurement error in our outcome variable and thus potentially bias our estimates. The results from omitting P1 in 2015 and 2016 are presented in table 6.

[Table 6 about here]

Table 6 shows that excluding all imputed P1 scores increases standard deviation of the within school teacher effects slightly to 0.19 SDs compared to 0.14 SDs in table 6. This suggests that our estimates in table 6 are slightly attenuated due to the imputation of the baseline scores.

6. Effects of the NULP

In Table 7, we show how our estimates are affected by the introduction of the NULP. In order to obtain a balanced sample across intervention groups when only using randomly assigned teachers we have to refrain from estimating teacher effects and instead estimate classroom effects using the same sample as in table 5 column 2[12].

[Table 7 about here]

Column one shows the results for the group of schools that did not get the program. We see that the within-school estimates (panel B) is quite close to the average classroom effect of 0.14 SDs (table 5 column 2)

The results in columns two and three reveal that the program greatly increases the variance of teacher effects. Since the program leads to gains in student performance on average, this suggests

[12] Restricting the sample to teacher effects reduced the number of teachers in the control group to 19 (8) teachers (schools); reduced cost group to 42(16) teachers (school) and full cost program to 58(20) teachers(schools). The results from estimating teacher effects can be seen in appendix D.

that the impact of the program was largest for the highest-quality teachers. In panel B, which presents the results from purging the school effect, we see that teachers receiving the full cost program increases student performance by 0.22 SDs. This in turn suggests an added benefit of improving teacher quality in low-resource African contexts: better teachers can make other interventions even more effective.

7. Who Are the Most Effective Teachers and What Do They Do?

Using data from the teacher surveys (available in 2013 and 2014) and classroom observations (available in 2013), we are able to describe what teacher characteristics and behaviors are correlated with higher value-added measures. As seen from table 8 we find no relationship between any of the teacher characteristics and our estimated teacher effects.

[Table 8 about here]

Moving to teacher behaviors we graphically present the correlation between our estimated teacher effects and various behaviors (see figure 1, 2 and 3).

[Figure 1 and 2 about here]

From figure 1 and 2 the most apparent difference between effective and ineffective teachers is that more effective teachers tend to have more planned lessens. In addition, more effective teachers also seem to more often encourage participation and observe and record performance compared to less effective teachers. Figure 3 shows that more effective teachers also tend to spend *less* time teaching in Leblango compared to less effective teachers. This result is surprising as the one of the core elements of the NULP is to support teachers in teaching in Leblango.

[Figure 3 about here]

However, splitting the sample by intervention group shows that this relationship is entirely driven by the control and reduced cost program schools (see figure 4) suggesting that it is the most effective teachers that are switching from teaching in English to teaching in Leblango as a result of the NULP.

[Figure 4 about here]

8. Conclusion

Despite severe problems with teaching quality we found that teachers do matter for student learning in Northern Uganda and that raising a teacher from the 10th to the 90th percentile causes a 0.36 SD improvement in student performance. In a recent meta-analysis based on 77 randomized experiments with a total of 111 treatment arms McEvan (2015) compared the effect of different school-based interventions on learning outcomes in developing countries. The mean effect size of the most effective programs lies between 0.035 and 0.15 standard deviations suggesting that improving the effectiveness of the worst teachers to that of the best teachers is a very effective way of improving learning outcomes.

Our estimate of teacher effectiveness of 0.14 standard deviations is slightly higher than that found for primary schools in the US 0.08 SDs (Chetty *et al.* 2014) and Ecuador 0.09 SDs (Araujo *et al.*2015). This suggests that teachers are at least as important in a low income context such as Uganda as it is in both high and middle income contexts.

In order to transform the knowledge of "teachers matter" to information that would be useful for policy makers and administrators to recruit, train and support teachers it is important to know who the most effective teachers are and what they do in the classroom. To address this issue we correlated our estimated teacher effects with teacher characteristics and classroom behaviors. We found no evidence of teacher characteristics being associated with teacher effectiveness and only suggestive evidence that more effective teachers have more planned lessons, encourage participation and observe and record performance. These findings (or lack of finding) are unfortunately common in the literature on teacher effectiveness (Araujo *et al.*, 2016; Azam & Kingdon, 2015; Slater *et al.*, 2012) and more research into who the most effective teachers are? and what behaviors make them effective? is most needed.

9. References

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y. & Schady, N. (2016): Teaching quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, in press.

Aaronson, D., Barrow, L. and Sander, W. (2007): Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics.* 25 no. 1, 95-135.

Azam, M. & Kingdon G. G. (2015): Assessing teachers in India. *Journal of Development Economics. 177,* 74-83.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of Abbreviated Nine-item Forms of the Raven's Standard Progressive Matrices Test. *Assessment*, 19(3), 354–369.

Bold, T., Filmer, D., Martin, G., Molina, E., Rockmore, C., Stacy, B., Svensson, J., Wane, W. (2017). What Do Teachers Know and Do? Does it Matter? Evidence from Primary Schools in Africa. *Policy Research Working Paper 7956.* World Bank Group.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011): How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *Quarterly Journal of Economics*, 126(4), 1593–1660.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014): Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104(9).* 2593-2632

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical

foundation, purpose, and limitations. *International Journal of Educational Development*.

Ganimian, A. & Murnane, R. (2014): Improving educational outcomes in developing countries: Lessons from rigorous impact evaluations. *NBER Working Paper no. 20284*. Cambridge, MA: National Bureau of Economic Research (NBER)

Glewwe, P. Hanushek, E., Humpage, S. D. & Ravina, R. (2014): School resources and educational outcomes in developing countries; A review of the literature from 1990 to 2010. In P. Glewwe (Ed.), *Education Policy in Developing Countries.* Chicago, IL and London UK: University of Chicago Press.

Gove, A., & Wetterberg, A. (2011). The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy.

Hanushek, E. A. & Rivkin, S. G. (2010): Generalizations about using value-added measures of teacher quality. *The American Economic Review, Vol. 100, No 2.* 267-271.

Hanushek, E. & Rivkin, S. (2012): The distribution of teacher quality and implications for policy. *Annual Review of Economics. 4.* 131-157

Hardman F., Ackers, J. , Abrishamian, N. & O'Sullivan, M. (2011) Developing a systemic approach to teacher education in sub-Saharan Africa: emerging lessons from Kenya, Tanzania and Uganda, Compare: A Journal of Comparative and International Education, 41:5, 669-683,

Horváth, H. (2015): Classroom Assignment Policies and Implications for Teacher Value-Added Estimation.

Kane, T. J. & Staiger, D. O. (2008): Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper no. 14607*. Cambridge, MA: National Bureau of Economic Research (NBER)

Kim, T., & Axelrod, S. (2005). Direct instruction: An educators' guide and a plea for action. *The Behavior Analyst Today*, 6(2), 111.

Kremer, M., Brannen, C. &Glennerster, R. (2013): The challenge of education and learning in the developing world. *Science, 340*, 297-300.

McEvan, P. (2014): Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, XX*, 1-42.

MoES (2013): Teacher issues in Uganda: A diagnosis for a shared vision on issues and the designing of a feasible, indigenous and effective teachers' policy. Ministry of Education and Sports.

Piper, B. (2010). *Uganda Early Grade Reading Assessment Findings Report: Literacy Acquisition and Mother Tongue*. Research Triangle Institute.

Rivkin, S., Hanushek, E. & Kain, J. (2005): Teachers, schools, and academic achievement. *Econometrica. 73(2).* 417-458

Rothstein, J. (2010): Teacher quality in educational production: Tracking, decay, and student achievement.*The Quarterly Journal of Economics, 125(1).* 175-214.

RTI International. (2009). *Early Grade Reading Assessment Toolkit*. World Bank Office of Human Development.

Slater, H., Davis, N. M. and Burgess, S. (2012): Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics*, *74(5).* 629-645

Todd, P. E. and Wolpin, K. I (2003): On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, Vol. 113, No. 485, Features (Feb., 2003), pp. F3-F33

Uwezo (2016): Are Our Children Learning (2016) , 2016. Twaweza.

http://www.uwezo.net/wp-content/uploads/2016/12/UwezoUganda2015ALAReport-FINAL-EN-web.pdf

Figures and Tables

Figure 1: Positive Teacher Behaviors

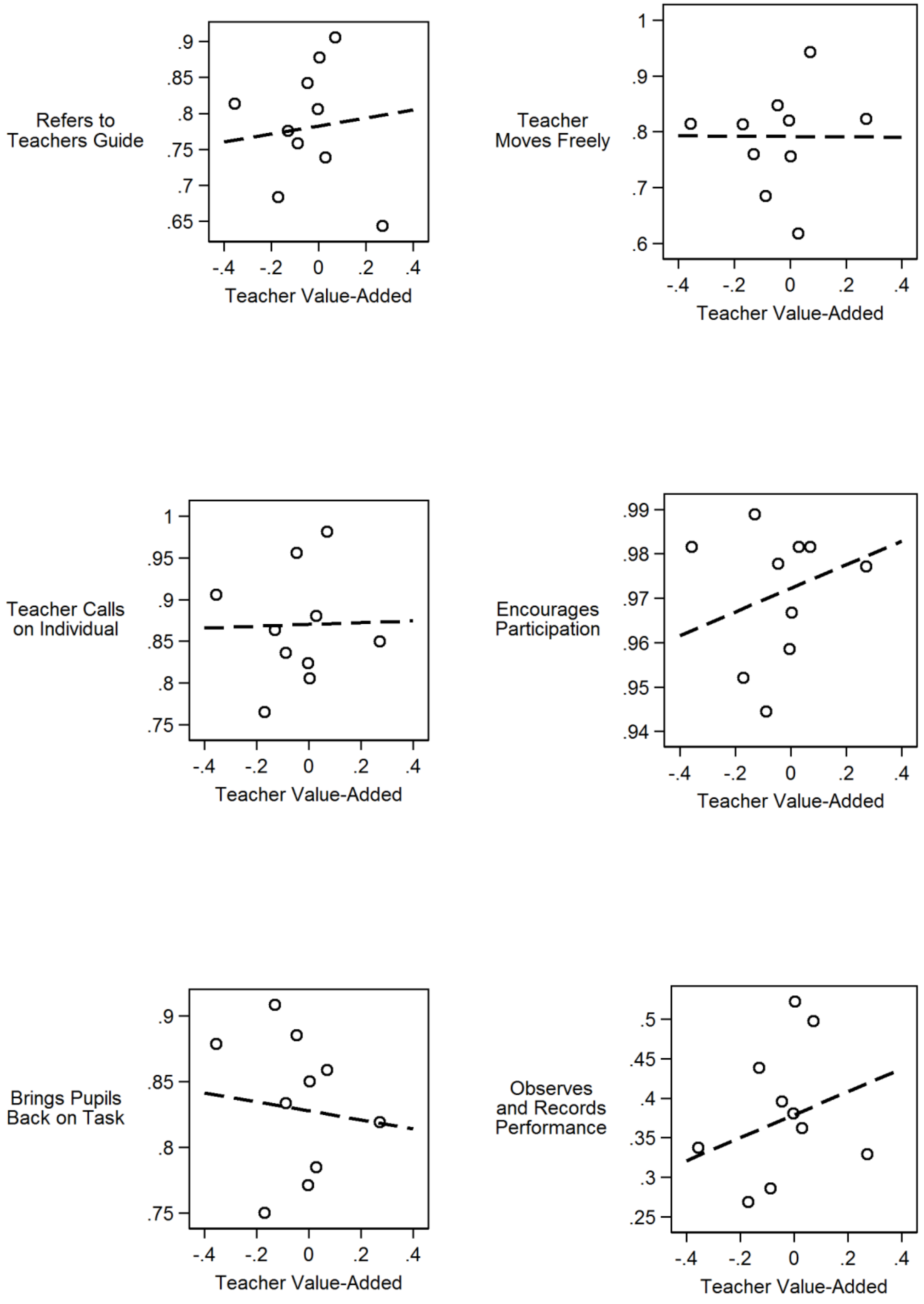


Figure 2: Negative Teacher Behaviors

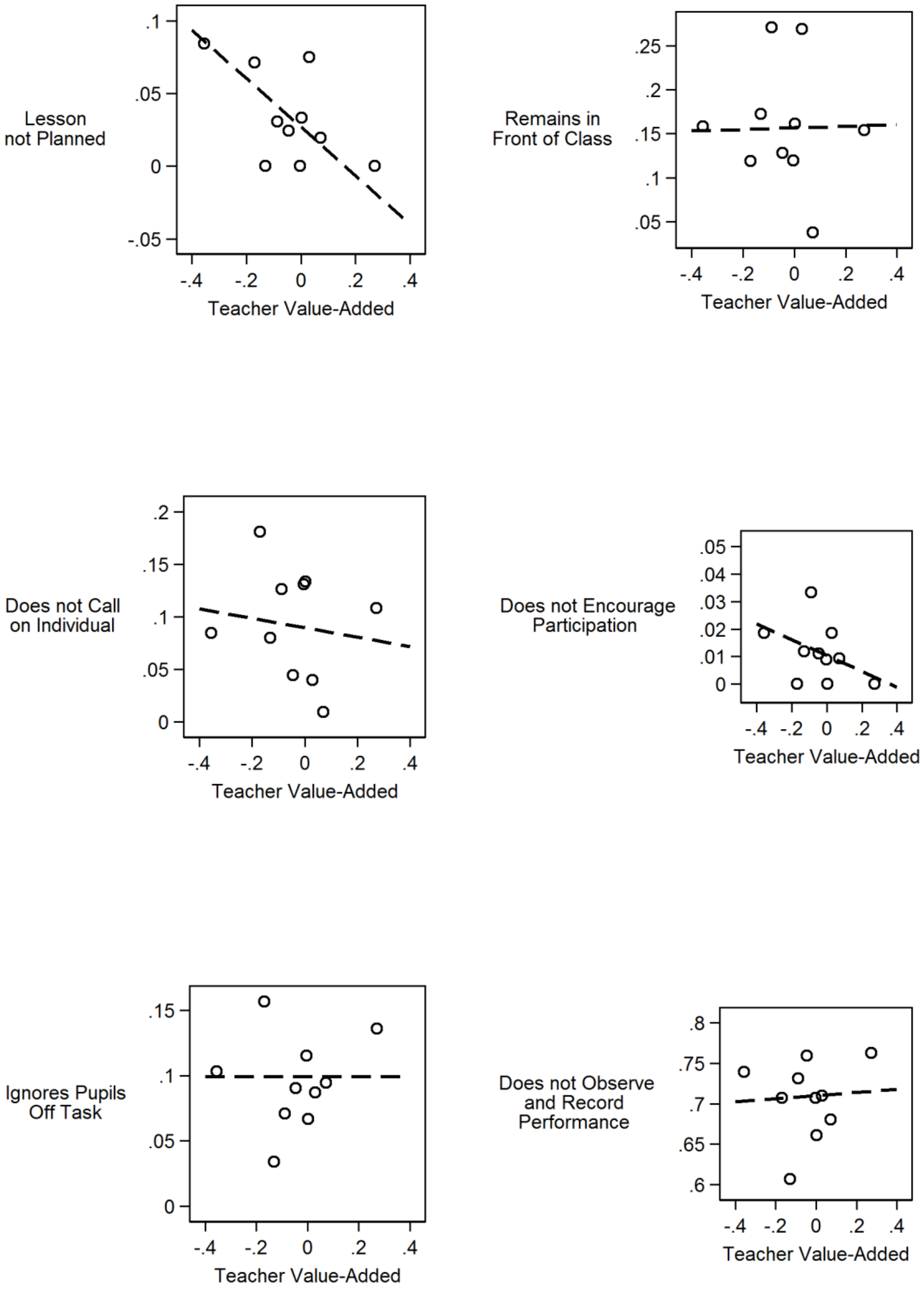


Figure 3: Instruction Language Behaviors

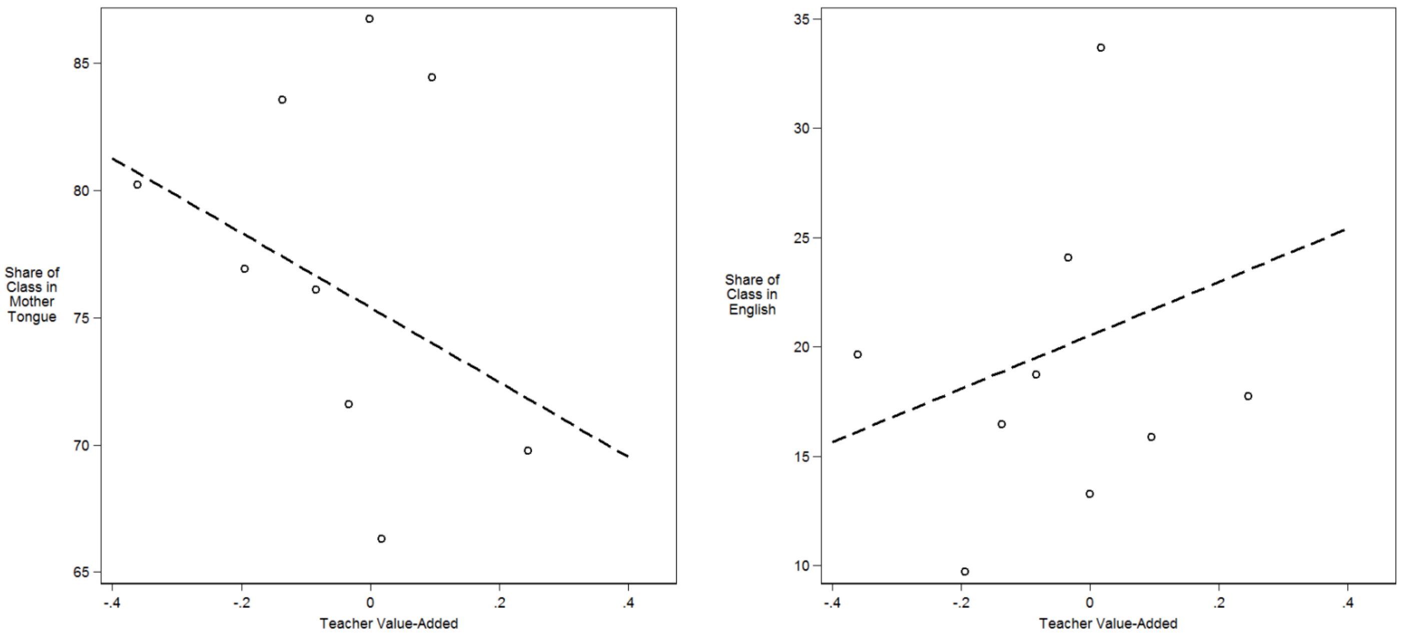


Figure 4: Use of Leblango by Intervention Group

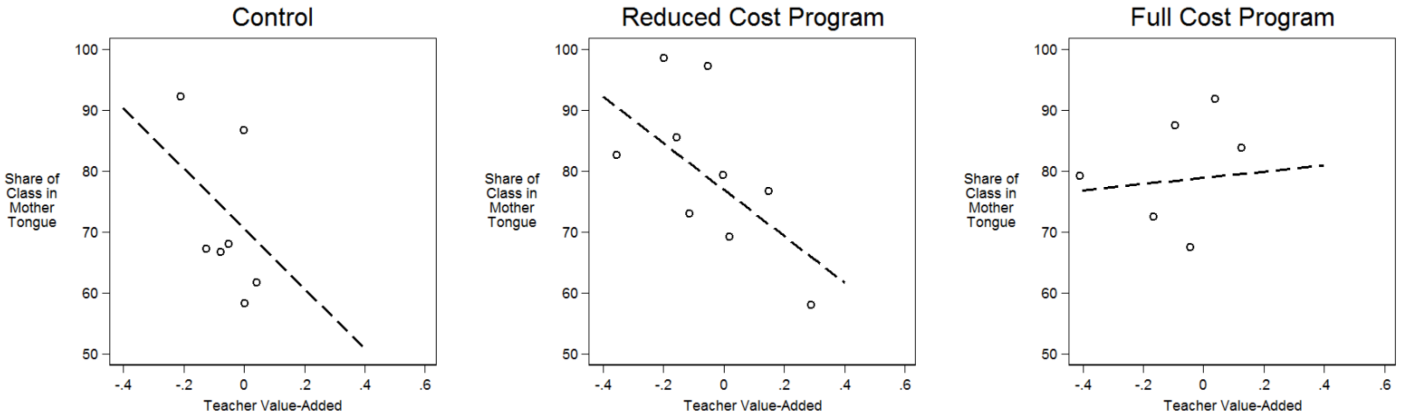


Table 1: Sample

	All teachers			Teachers with at least two years of data		
Panel A: Students	Control	Reduced-cost program	Full-cost program	Control	Reduced-cost program	Full-cost program
# Children	6387	6997	6805	4496	5734	5396
# Children x year	9498	10485	10387	5227	6759	6738
Panel B: Teachers						
# Teachers	235	240	237	77	102	100
Panel C: Schools						
# Schools	42	44	42	40	44	41

Table 2: Descriptive statistics

	All teachers			Teachers with at least two years of data		
	Control	Reduced-cost program	Full-cost program	Control	Reduced-cost program	Full-cost program
P1						
# Children	5094	5589	5280	3769	4550	4282
Age	7.53	7.57	7.54	7.55	7.60	7.54
Female (%)	0.49	0.51	0.50	0.48	0.51	0.49
Baseline score	0.00	-0.01	0.01	0.01	0.00	0.01
Endline score	0.32	0.55	0.75	0.33	0.55	0.76
P2						
# Children	2401	2633	2779	1178	1672	1820
Age	8.80	8.83	8.84	8.79	8.79	8.81
Female (%)	0.49	0.50	0.50	0.46	0.51	0.50
Baseline score	0.01	0.21	0.43	-0.02	0.27	0.47
Endline score	0.24	0.82	1.35	0.23	0.93	1.40
P3						
# Children	1776	1917	2050	249	525	624
Age	9.82	9.85	9.83	9.52	9.76	9.75
Female (%)	0.50	0.50	0.49	0.50	0.52	0.52
Baseline score	0.02	0.38	0.73	0.13	0.14	0.57
Endline score	0.41	1.05	1.68	0.50	0.55	1.47
P4						
# Children	227	346	278	31	12	12
Age	10.80	10.90	10.69	10.84	10.50	10.50
Female (%)	0.52	0.48	0.46	0.42	0.58	0.42
Baseline score	0.15	0.11	0.38	-0.02	-0.30	-0.24
Endline score	0.50	0.68	0.86	0.37	0.26	0.78

Table 3: Baseline Results

	Classroom Effects	**Classroom Effects**	**Teacher Effects**
Sample:	*Full sample*	Multiple years	*Multiple years*
Panel A: All teachers			
SD	0.56	0.58	0.49
Corrected SD	0.51	0.52	0.43
Children	30094	18342	18342
Teachers	714	275	275
Schools	128	125	125
Pupils per teacher	28	28	32
Panel B: School FE			
SD	0.39	0.39	0.25
Corrected SD	0.33	0.31	0.19
Children	27111	12939	12939
Teachers	688	248	248
Schools	127	98	98
Pupils per teacher	27	27	30

Table 4: Bias

	(1)	(2)
Classroom effects	0.680	0.472
(non-random assignment)	(0.018)***	(0.032)***
T-test ($\hat{\lambda}_{cs} == 1$)	0.000	0.000
School FE	NO	YES

Table 5: Results Using Only Randomly Assigned Teachers

	(1) **Classroom Effects**	**(2)** **Classroom Effects**	**(3)** **Teacher Effects**
Sample	*Only 2013 and 2016*	*All years where random assignment cannot be rejected*	*As (2) plus teacher present multiple years*
Panel A: All teachers			
SD	0.40	0.47	0.42
Corrected SD	0.34	0.41	0.34
Children	14920	20145	8964
Teachers	501	603	156
Schools	128	128	81
Pupils per teacher	29	27	30
Panel B: School FE			
SD	0.25	0.26	0.23
Corrected SD	0.16	0.14	0.14
Children	14379	18811	5936
Teachers	496	589	119
Schools	127	127	44
Pupils per teacher	28	27	28

Table 6: Robustness, excluding all P1 pupils in 2015 and 2016

	Classroom Effects	Teacher Effects
Panel A: All teachers		
SD	0.55	0.54
Corrected SD	0.47	0.46
Children	5023	5023
Teachers	142	142
Schools	77	77
Pupils per teacher	21	22
Panel B: School FE		
SD	0.33	0.29
Corrected SD	0.20	0.20
Children	2638	2638
Teachers	82	82
Schools	31	31
Pupils per teacher	21	22

Table 7: Heterogeneity of the classroom effects by treatment group

Panel A: All teachers	**Control**	**Reduced-cost program**	**Full-cost program**
SD of classroom effects	0.34	0.45	0.50
Corrected SD of classroom effects	0.28	0.38	0.43
Children	5991	7059	7095
Teachers	197	206	200
Schools	42	44	42
Panel B: School FE			
SD of classroom effects	0.21	0.25	0.31
Corrected SD of classroom effects	0.13	0.15	0.22
Children	5553	6531	6727
Teachers	188	203	198
Schools	41	44	42

Table 8: Relationship between teacher effectiveness and teacher characteristics

Dependent variable: *Teacher effect*	
Age	0.001
	(0.002)
Years of schooling	0.002
	(0.011)
Ravens Progressive Matrices	0.010
	(0.011)
Salary	0.000
	(0.000)
Gender	-0.011
	(0.030)
Observations	115
R-squared	0.029

Appendices

Appendix A: Principal Component Analysis

2013

Appendix Table 1: Results of Principal Component Analysis P1 2013

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	3.21	2.33	0.53	0.53
Second	0.87	0.07	0.15	0.68
Third	0.80	0.22	0.13	0.81
Fourth	0.59	0.25	0.10	0.91
Fifth	0.34	0.15	0.06	0.97
Sixth	0.19		0.03	1.00

2014

Appendix Table 2: Results of Principal Component Analysis P1 2014

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	2.97	1.96	0.50	0.50
Second	1.02	0.11	0.17	0.67
Third	0.91	0.22	0.15	0.82
Fourth	0.68	0.38	0.11	0.93
Fifth	0.30	0.18	0.05	0.98
Sixth	0.12	.	0.02	1.00

2015

Appendix Table 3: Results of Principal Component Analysis P1 2015

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	2.89	1.88	0.48	0.48
Second	1.01	0.07	0.17	0.65
Third	0.94	0.35	0.16	0.81
Fourth	0.58	0.26	0.10	0.90
Fifth	0.33	0.08	0.05	0.96
Sixth	0.25	.	0.04	1.00

Appendix Table 4: Results of Principal Component Analysis P2 2015

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	3.38	2.48	0.56	0.56
Second	0.90	0.05	0.15	0.71
Third	0.85	0.42	0.14	0.86
Fourth	0.43	0.16	0.07	0.93
Fifth	0.27	0.10	0.04	0.97
Sixth	0.17	.	0.03	1.00

Appendix Table 5: Results of Principal Component Analysis P3 2015

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	3.95	3.13	0.66	0.66
Second	0.82	0.18	0.14	0.79
Third	0.64	0.29	0.11	0.90
Fourth	0.35	0.22	0.06	0.96
Fifth	0.13	0.03	0.02	0.98
Sixth	0.10	.	0.02	1.00

2016

Appendix Table 6: Results of Principal Component Analysis P1 2016

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	2.65	1.69	0.44	0.44
Second	0.95	0.08	0.16	0.60
Third	0.87	0.12	0.15	0.75
Fourth	0.76	0.22	0.13	0.87
Fifth	0.53	0.30	0.09	0.96
Sixth	0.24	.	0.04	1.00

Appendix Table 7: Results of Principal Component Analysis P2 2016

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	3.28	2.29	0.55	0.55
Second	0.99	0.27	0.17	0.71
Third	0.73	0.25	0.12	0.83
Fourth	0.48	0.17	0.08	0.91
Fifth	0.31	0.10	0.05	0.96
Sixth	0.21	.	0.04	1.00

Appendix Table 8: Results of Principal Component Analysis P3 2016

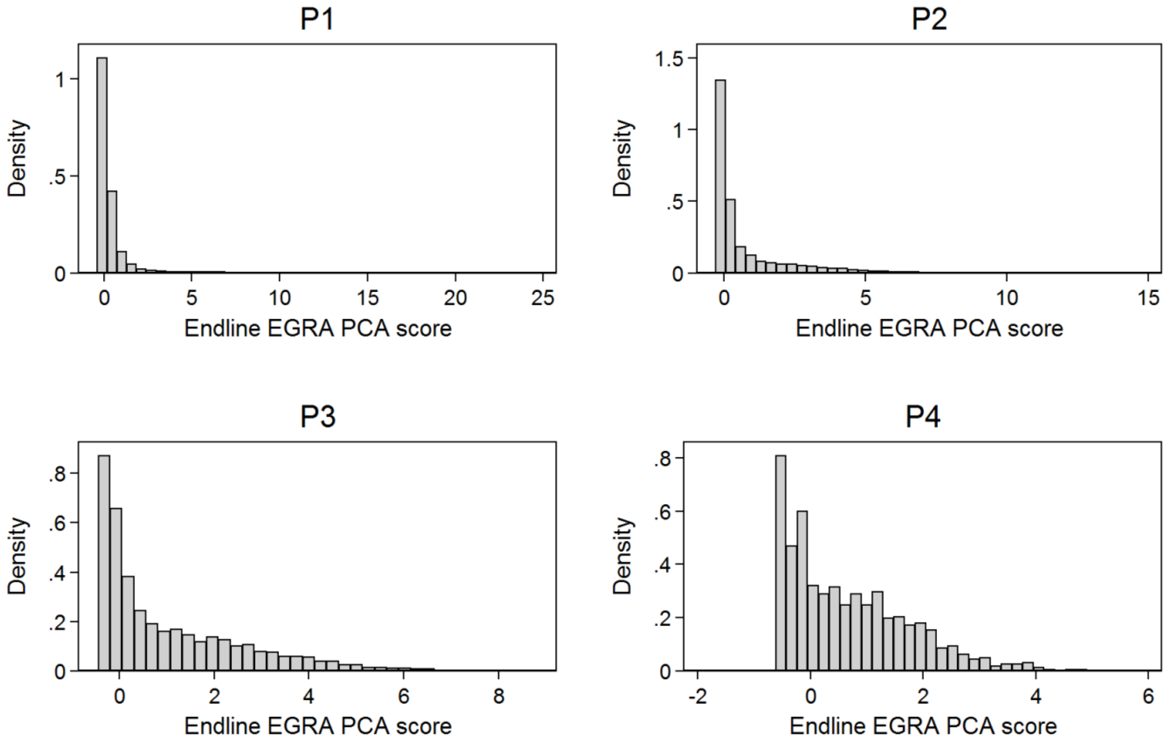
	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	3.95	3.13	0.66	0.66
Second	0.82	0.23	0.14	0.80
Third	0.59	0.20	0.10	0.89
Fourth	0.39	0.25	0.07	0.96
Fifth	0.14	0.04	0.02	0.98
Sixth	0.11	.	0.02	1.00

Appendix Table 9: Results of Principal Component Analysis P4 2016

	Eigenvalue	Difference from Next-Largest Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained
Component				
First	3.95	3.11	0.66	0.66
Second	0.84	0.16	0.14	0.80
Third	0.68	0.32	0.11	0.91
Fourth	0.36	0.25	0.06	0.97
Fifth	0.11	0.05	0.02	0.99
Sixth	0.07	.	0.01	1.00

Appendix B: Distributions of Endline PCA Scores by Grade Level

Figure B1: Distributions of Endline PCA Scores by Grade Level



Appendix C Distributions of Baseline Subtests for P1 in 2013 and 2014

Figure C1: Distribution of the raw scores in the subtest for P1 in 2013

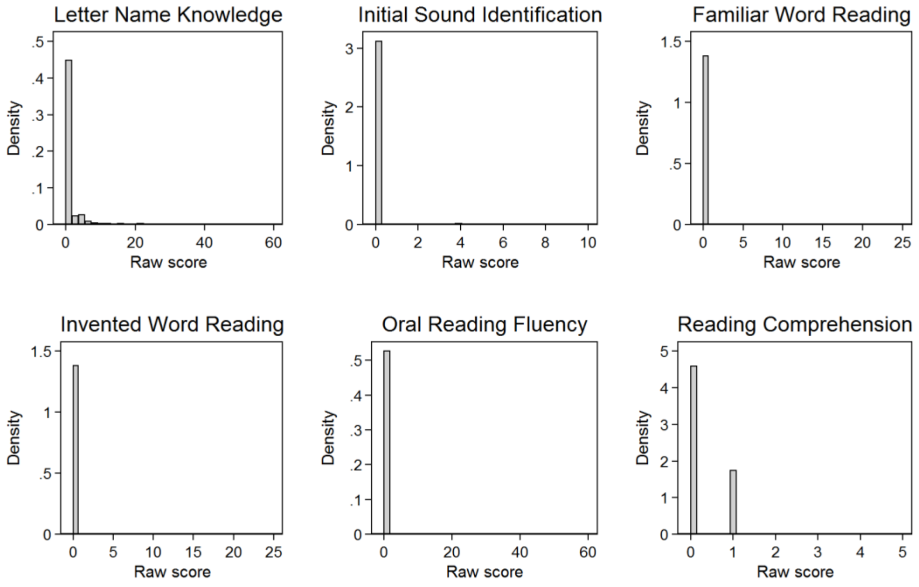
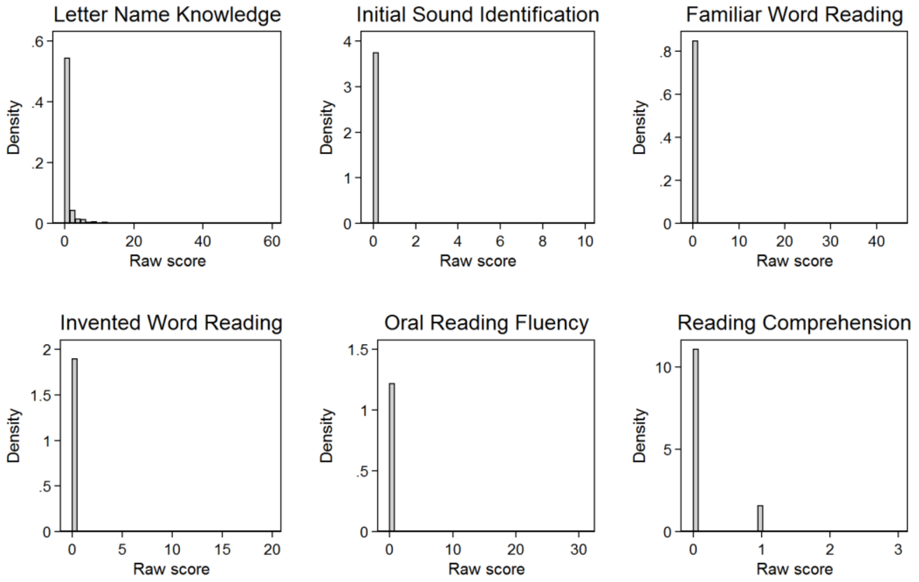


Figure C2: Distribution of the raw scores in the subtest for P1 in 2014



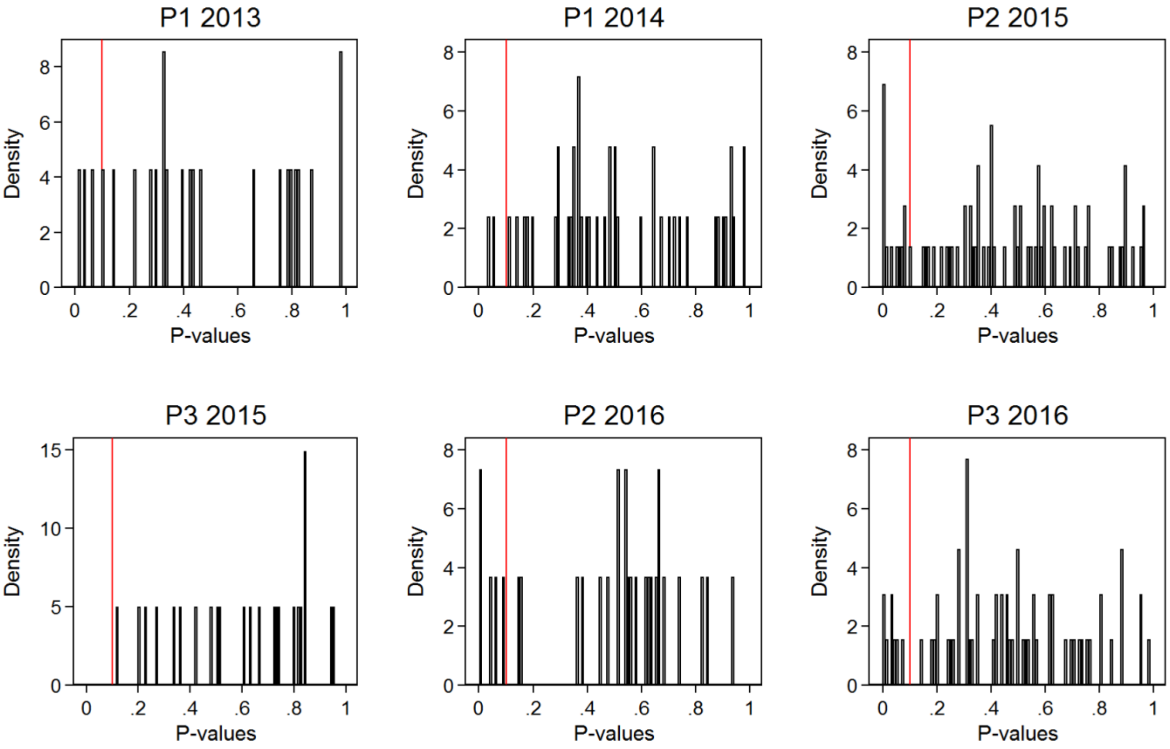
Appendix D: Heterogeneous Teacher Effects

Appendix table D1 : Heterogeneity of the teacher effects by treatment group

Panel A: All teachers	Control	Reduced-cost program	Full-cost program
SD of classroom effects	0.31	0.37	0.44
Corrected SD of classroom effects	0.22	0.29	0.35
Children	1705	3378	3881
Teachers	30	58	68
Schools	19	32	30
Panel B: School FE			
SD of classroom effects	0.17	0.23	0.25
Corrected SD of classroom effects	0.04	0.14	0.15
Children	912	2116	2908
Teachers	19	42	58
Schools	8	16	20

Appendix E Verifying Random Assignment in 2013 and 2016

Figure E1: Distributions of P-values testing differences in baseline scores between classrooms within each school



Notes: The red line marks a P-value of 0.1