Teacher Effectiveness in Africa: Longitudinal and Causal Estimates

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

March 4, 2019

[*Draft: Please no not quote or cite without permission*]

Abstract

This paper presents the first estimates of teacher effectiveness from Africa using longitudinal 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 the variation in teacher effectiveness. A 1-SD increase in teacher effectiveness leads to at least a 0.13 SD improvement in student reading at the end of one year. Using detailed survey and classroom observation data, we find no detectable correlation between teacher effectiveness and teacher characteristics, but do find patterns associated with teaching behavior in the classroom. Using the RCT we find that providing teacher training and support increases the variation in teacher effectiveness, by probably making the most-effective teachers relatively better than the least-effective teachers.

[1] Buhl-Wiggers: Department of Economics, Copenhagen Business School (jubu.eco@cbs.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 Wisconsin (econjeff@ssc.wisc.edu).

Acknowledgements: We thank Laura Schechter and Chao Fu and seminar audiences at the University of Minnesota, CSAE, RISE, the University of Wisconsin, SOLE, Stellenbosch University, International Population Conference, NYU Steinhardt and Northeastern University for their comments and suggestions. The randomized evaluation of the Northern Uganda Literacy Project would not have been possible without the collaboration of Victoria Brown, Katherine Pollman, Deborah Amuka and other Mango Tree Educational Enterprises staff.

1. Introduction

There are two main bodies of literature in economics that focus on understanding the relationship between teachers and student learning. The first uses student test scores to estimate teacher value added: extensive evidence from developed countries shows that exposure to teachers with higher value added scores has large effects on children's success in school and in adulthood (see eg. Rivkin, Hanushek, and Kain 2005, Chetty et al. 2011, Chetty, Friedman, and Rockoff 2014). A second body of literature compares the results from educational program evaluations – primarily conducted in developing countries – and finds that interventions that support and train teachers or focus on teaching methods and pedagogy, are the most effective at improving student learning (see e.g. Glewwe and Muralidharan 2016, Kremer, Brannen, and Glennerster 2013, McEwan 2015, Ganimian and Murnane 2014, Evans and Popova 2016). To date, these literatures have accumulated evidence in mainly separate spheres: value added studies conducted in developed countries and randomized control trials conducted in developing countries. This paper integrates these two approaches to shed light on the relationship between teachers and student learning in Uganda.

The specific aims of this study are threefold. First, we present the first value-added estimates of teacher effectiveness from an African country; our results are among the first from any developing country. We compare estimated classroom effects to teacher effects, and compare estimates when students are randomized to classrooms with when they are not. Second, to understand who effective teachers are, and what they do, we correlate our estimated teacher effects with teacher characteristics and classroom observation data. Third, we estimate the impact of a randomized intervention of a comprehensive teacher training and pedagogy program on the variation in teacher effectiveness. Contrary to previous literature, we are able to test how an effective teacher training and pedagogy program affects teacher value-added.

We use panel data from a randomized evaluation of a mother-tongue literacy program implemented in grades one to four in northern Uganda – the Northern Uganda Literacy Program (NULP) – to estimate teacher effectiveness. The program provided primary schools with intensive teacher training and support, scripted lesson plans, and revised learning materials. The program was developed by an Ugandan-based educational tools company – Mango Tree. Mango Tree began the program in a small number of pilot schools in 2010, where the materials and delivery of the program were 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 128 schools. In the evaluation, schools were assigned to one of three study arms: 1) full-cost, 2) reduced-cost, and 3) control. In the full-cost group, schools received the NULP program delivered by Mango Tree and its staff. In the reduced-cost group, some of the materials were eliminated, teacher training and support was conducted through a cascade model in collaboration with government tutors, and teachers received fewer support visits. The analysis of the effects of the program suggests massive effects on student learning – a 1.35 standard deviation increase in reading test scores for the full program and 0.78 in the reduced-cost version, after three years of the intervention (Buhl-Wiggers, et al. 2018).

In this paper, we utilize two aspects of this program. First, students were randomly assigned to classrooms within both treatment and control schools in 2013, 2016 and 2017 enabling us to address the issue of bias due to sorting of students to teachers when estimating teacher effectiveness (Chetty et al. 2014, Koedel and Betts 2011, Rothstein 2009). Second, using the randomization of the NULP across schools, we estimate the causal impact of teacher training on the variance of teacher effectiveness. This provides insight into whether teacher training and support make teachers more similar or more varied in their ability to affect student learning, and allows us to investigate which types of teachers are most likely to see gains in value-added. These insights add to the understanding and interpretation of teacher value-added estimates as we test what actually happens to the variation in teacher value-added when exposed to an effective teacher training program.

Our lower-bound estimate of the teacher value-added is that a one-standard deviation increase in teacher effectiveness improves local language reading test scores by 0.13 standard deviations. These lower-bound estimates are derived from within-school variation, corrected for sampling variation and strikingly similar to comparable estimates in other contexts. For example, the estimated effect of a one standard deviation increase in within school teacher effectiveness from schools in the United States, varies from 0.08 to 0.26 standard deviations of test scores (Hanushek and Rivkin 2010). Comparing our estimates to studies in low-resource settings is difficult because studies estimating teacher effectiveness in developing countries are scarce. In Ecuador, Araujo et al. (2016) find that a one standard deviation increase in within school teacher effectiveness increases test scores by 0.09 standard deviations among kindergarteners. In Pakistan, Bau and Das (2017) find that a one standard deviation increase in within school teacher effectiveness increases student performance by 0.16 standard deviations. Among private secondary school teachers in India, Azam and Kingdon (2015) find that a one standard deviation improvement in within school teacher effectiveness increased test scores by 0.37 standard deviations (over two years).[2]

Linking teacher effectiveness to teacher characteristics we find that more effective teachers also have more years of education and that less effective teachers tend to have higher salaries., Using a rich set of classroom observations data, we find limited associations between teacher effectiveness and teacher or student behaviors in the classroom.

When we evaluate the effects of the NULP intervention, we find an increase in the spread of classroom value-added. Compared to the control group estimate of 0.07 standard deviations, one standard deviation increase in teacher effectiveness in full-cost program schools leads to an increase in student performance of 0.15 standard deviations. We find some evidence that the largest gains in value-added are among the best teachers.

Direct evidence on the effects of teaching quality in Africa is scant. Such evidence is needed: if variation in teaching quality drives large changes in student performance, there is scope for policymakers and administrators to improve learning by either emulating the training of the most effective teachers, providing quality teacher support and mentoring or selective removal of the worst performing teachers.

Our findings have several implications. First, even in a low-resource context, teachers are important for student learning. Second, better teachers appear to gain more from teacher training and support, making it crucial to better understand how to reach teachers who are struggling to perform.

2. Setting and Intervention Details

2.1 Primary Education in Uganda

[2] A related literature examins the value-added of schools rather than teachers. We are aware of three papers that study school value-added in developing countries: Crawfurd and Elks (2018), for Uganda, Blackmon (2017), for Tanzania, and Muñoz-Chereau and Thomas (2016), for Chile.

Primary education in Uganda consists of seven years of schooling starting at age six. The vast majority of Ugandan children have attended school at some point in time and the net enrollment rate is above 90% (World Bank 2013). Despite this relatively high level of access, late enrollment, repetition and early drop out remain major challenges throughout the country. Only about 60% of students transition from primary to secondary school (World Bank 2010).

Since 1997, primary school has officially been free of charge, however, as resources are scarce many schools still depend on contributions from parents. The reform of 1997 was successful in getting children into school (Deininger 2003). 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 primary school curriculum. This new curriculum induced two main changes: Shifting the language of instruction from English to the local language (11 different languages of instruction throughout the country) in lower primary (grades 1 to 3) and implementing a thematic curriculum instead of the traditional subject-based curriculum.

Despite these changes, Uganda still faces major learning challenges in its primary schools. Bold et al. (2017) find that the vast majority (94%) of children in government primary schools could not read a 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. Even at the end of primary school, students have often learned very little: 15% of all grade 7 students leave primary school without mastering division and 20% leave primary school without being able to read a short story (Uwezo 2016). The figures for grade 7 likely overstate student performance, because schools discourage weaker students from attending in grade 7 in order to focus on preparing the strongest students for the higher-stakes primary leaving exam (Gilligan et al. 2018).

2.2 Teachers in Uganda

Primary school teachers must have obtained a Grade III Teacher Certificate to teach in Uganda. This requires four years of secondary school (O-level) followed by two years of pre-service teacher training. In 2010, the Ugandan Ministry of Education and Sports found that 12.7% of primary school teachers did not have the correct qualifications to teach. Yet even among qualified teachers, weaknesses in classroom pedagogy are 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 across Africa, Bold et al. (2017) find that 16% have minimum knowledge in language, 70% have minimum knowledge in math and only 4% have minimum pedagogical knowledge. In regards to classroom practices, most teachers give positive feedback, but only half or less ask a mix of lower and higher order questions. Similarly low shares of teachers plan their lessons in advance, or introduce and summarize their lessons. Very few teachers (5%) engage in all of the above practices.

These weaknesses have led to a larger focus on in-service education and especially Continuous Professional Development (CPD) which systematically updates competences that teachers require in the classroom. The CPD program is coordinated by the primary teachers' colleges through Coordinating Center Tutors (CCTs). CCTs 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 mentor other teachers (Hardman et al. 2011).

In addition to poor knowledge and pedagogical skills low levels of effective teaching time is also 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 the classroom leading to almost half of the classrooms being without a teacher (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 are then sent to the central government. As teachers are scarce, the first step is to re-allocate teachers from schools with a surplus of teachers to schools with a lack of teachers within the same district. When this is done the total amount of teachers that can feasibly 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 (Ugandan Ministry of Education and Sports 2014).

Teacher attrition from teaching is estimated to be around 4% annually 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 (Ugandan Ministry of Education and Sports 2014). 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 an early grade mother-tongue literacy program developed in response to the educational challenges facing northern Uganda. The NULP was designed by a locally owned educational tools company, Mango Tree, and is based in the Lango sub-Region, where the vast majority of the population speaks one language – Leblango. The NULP involves providing residential teacher training throughout the school year and classroom support visits to give feedback to teachers. The program's pedagogy involves training teachers how to be more engaged with students, and moving through material at a slower pace to ensure the acquisition of fundamental literacy skills. Teachers are provided with detailed, scripted guides that lay out daily and weekly lesson plans, as well as new primers and readers for every student, and slates, chalk, and wall clocks for first-grade classrooms.[3]

The NULP was introduced to different grades during the time of our study. In 2013 and 2014, all first-grade classrooms and teachers received the NULP, in 2015 second-grade classrooms and teachers received the program, and 2016, all third-grade teachers received the program.[4]

[3] 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.

[4] In 2017, Mango Tree piloted a teacher mentor program with fourth-grade teachers to provide support, but no materials or pedagogical trainings or support were delivered.

Classrooms were allowed to keep all of the Mango Tree educational materials (such as slates, primers, and readers) after they received the program, but teachers were no longer provided additional training or support visits. If new teachers were transferred into a classroom that had previously received the NULP, they were also not give additional training or support.

3. Sample, and Data

3.1 Sample

Schools

There are a total of 128 schools in our study. Schools were sampled for the study in two phases. 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 grade-one classrooms and teachers.[5] In 2014 the program was expanded to 90 additional schools for a total of 128 schools. The eligibility criteria for these new schools were slightly different, and less stringent.[6] The number of classrooms per grade was no longer stipulated.

Students

We use data collected over five years 2013-2017, consisting of four cohorts of grade-one children who entered the study schools in 2013, 2014, 2015, and 2016. Depending on the cohort, we follow the students from grade one to either grade two, three, four or five. Our sample of teachers corresponds to the classrooms that are studied from the four cohorts of students.

In 2013, 50 grade-one students were randomly sampled 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, 2016 and 2017 this initial sample of grade-one students was retained, and tracked into grades two, three, four, and five. In 2014, a new cohort of grade-one students was added to the study. Among this new cohort,

[5] Other eligibility criteria include: being located in one of five specific school districts (coordinating centres), 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.

[6] Criteria in 2104 include: having desks and blackboards in 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.

100 grade-one students were randomly selected from each of the 128 schools.[7] As with the first cohort, this cohort was also tracked into grades two, three, and four in 2015, 2016, and 2017 respectively. In 2015, a third and smaller cohort, 30 grade-one students randomly selected from each school, was added and tracked into grades two and three in 2016 and 2017. In 2016, the fourth cohort was added, by randomly sampling 60 grade-one students in each school.

3.2 Randomization

Assignment of students to classrooms and teachers

In three of the five years considered, 2013, 2016 and 2017 students were randomly assigned to classrooms. To do so, we provided head teachers in each school with blank student 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 who 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 interviewing head teachers. In order to test compliance, we compare baseline score means across classrooms within schools and grade level each year[8]. We find that between 1 to 5% of the schools had classrooms with statistically significant baseline differences between streams at the 5%-level.[9] We do robustness checks excluding those classrooms.

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. From head teacher surveys we have information that approximately 60 percent of the surveyed head teachers do not sort students on ability, behavior, gender, parental preferences nor friends. For the remaining head teachers some

[7] 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 grade one pupils was drawn at baseline 2014, and then 60 students were added at endline 2014 following the same sampling procedure as at baseline. In the 90 new schools, the initial sample was 80 pupils and 20 additional pupils were added at endline. The difference in sampling strategy was due to the organizational difficulty of handling large numbers of students to test at baseline or endline.

[8] This is a similar to the approach suggested by Horvath (2015).

[9] 5.3 % in P1, 0.8% in P2, 3.1% in P3, 3.1% in P4 and 0.8% in P5. See Appendix A for distributions of the P-values.

general patterns are evident. First, head teachers attempt to balance the gender of students across classrooms. Second, the few head teachers (less than one percent of the surveyed teachers) that sort on ability or behavior assign better students to better teachers and worse behaved students to worse teachers. Given these answers by the head teachers, it is plausible that in the business-as-usual years (2014 and 2015) there is little systematic sorting of students to classrooms on the basis of ability. When testing differences in baseline test scores between classrooms in these years we find similar numbers as in the random assignment years, namely that 0 to 5% of the schools had classrooms with statistically significant baseline differences between streams.[10]

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 and Thornton (2017)). 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 Ministry of Education coordinating center tutors (CCTs) rather than Mango Tree staff, and teachers received fewer support visits, from CCTs. 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.

3.3 Analytical Sample Construction

We start with 58,782 students-year observations across all grades and years. Of these, we are able to match 58,231 to class teachers, through student and teacher reporting's. In order to limit estimation error due to misreporting we delete observations where we have less than five students per teacher and thus drop 1,909 observations. Moreover, as we need baseline test scores to estimate the value-added model we also restrict the sample to only including children with baseline scores (dropping 13,795 observations). Our final restriction is that we need at least two teachers in each

[10] The exact numbers are 2.3% in P1, 5.0% in P2 and 0% in P3

school each year to purge the school effect. This leads us to a total sample of 39,911 student-year observations (23,105 unique students) and 1,672 teacher-year observations (1,085 teachers unique teachers).

Teachers

We work with four main samples of teachers. Table 1 presents the sample statistics for each of the analytical samples. The *full sample* includes all teachers available from the study, and is used to estimate classroom effects. The *longitudinal sample* restricts the sample to teachers who are in the data across multiple years, as this is needed in order to estimate teacher effects. The *random assignment sample* is a subsample of the full sample, restricted to years where students were randomly assigned to teachers (2013, 2016 and 2017). The *longitudinal random assignment sample* further restrict the random assignment sample to only include teachers that teach in two of the three random assignment years.

[Table 1 about here]

Our sample of teachers is largely grade-specific rather than cohort-specific, 92% of the teachers teach a specific cohort once and then move on to a new cohort. In total, we have 1,085 teachers across all years and grades; of these 447 (or 41%) we observe teaching in at least two years. Thus, despite the fact that we have data from four adjacent years there is a relatively high change in the body of teachers, as over half of teachers are only observed in one year.

3.4 Measures and Descriptive Statistics

Summary statistics for students and teachers are presented in Table 2. For students, the average age across years and grades is around 9 years and approximately 50 percent are girls. The baseline score of each student is defined as the test score in the previous year i.e. if a student is in grade 3 the baseline score is the test score from grade 2. As we have a longitudinal RCT it means that baseline scores in grades 3 to 5 are measured post treatment and thus should not be interpreted as pre-treatment baseline scores. From Table 2 we also see clear treatment effects in the endline scores.

[Table 2 about here]

Learning Outcomes

Our primary outcome of interest comes from the Early Grade Reading Assessment (EGRA), an internationally recognized exam to assess early literacy skills such as recognizing letters, reading simple words and understanding sentences and paragraphs (Dubeck and Gove 2015, Gove and Wetterberg 2011, RTI 2009, Piper 2010). We use a validated adaptation of the EGRA to the local language (Leblango). The test 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 performance we construct an index in the following way. First, we calculate the mean of the test modules for each student-year-grade observation. Second, we standardize that against the control group separately for each year and grade.[11]

Tests were administered at the beginning and end of the year in both 2013 and 2014. In 2015, 2016 and 2017 the tests were only administered at the end of the year. Because the vast majority of grade-one students (90%) score zero across all questions and subtasks when tested at baseline in 2014 we find it reasonable to set the baseline score for grade one in 2015 and 2016 to zero.[12] This means that for all grade-one students, the value-added is from no skill to the skills obtained at the end of the year.

Teacher Characteristics and Teaching Practices

Data on teacher characteristics are obtained from teacher surveys conducted in 2013, 2014, 2015, and 2017. From these surveys we have information on both individual and household

[11] All results are robust to using a principal components score index (as in Black and Smith (2006)) instead. Results available upon request.

[12] See Appendix B for the distributions of the baseline subtest in 2013 and 2014.

characteristics.Of the 1,085 teachers in our sample, we have characteristics for approximately half (665 teachers). Table 2 shows that the average teacher is around 43 years old, has 14 years of education (which corresponds to two years of post-secondary education), 16 years of teaching experience and earns 405,000 shillings per month[13] ($105). Roughly 43 percent are women.

In 2013 and 2017 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 actions and behavior, the use of Leblango and English, and time spent on various teaching activities. Observations were conducted three times that year, in July, August and October. Each 30-minute lesson was broken up into three 10-minute observation blocks; for each block of time, the enumerator ticked off boxes to indicate which of the specified actions which occurred.

Following Glewwe, Ross, and Wydick (2017) we conduct a factor analysis to summarize the classroom observations into broader categories of behaviors. We retain all factors that explain at least 10% of the variance in the data and then apply a varimax rotation to the resulting set of selected factors (see Kerwin and Thornton (2018)). We estimate three factors from nine different teacher actions: "Keep Students Focused", comprising of bringing students back on task and not ignoring off-task students, "Solid Lesson Plan" comprising referring to a teacher's guide, participating, and having a planned lesson, and "Active Throughout Classroom", comprising moving freely around the classroom, calling on individuals, and observing student performance.

We also use data from the observations that occurred either during reading or writing activities. In particular, we look at the elements of focus of the lesson (sounds, letters, words, or sentences for reading and pictures, letters, words, sentences and name for writing), the percent of pupils participating, the materials used during the lesson (board, primer, or reader for reading and board, slate, or paper for writing), and teaching approach during the lesson (whole class, small group, individual at seat, or individual at the board for reading, and writing with motions in the air, practicing handwriting, copying text from the board, and writing ones own text for writing). We also observe the participation of students during speaking and listening activities (ie. not on the board and not using printed text) and whether they are working with a partner, small group, entire class, or with the teacher.

[13] Varying from 100,000 to 960,000 shillings per month.

4. Conceptual Framework and Empirical Strategy

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. Todd and Wolpin (2003) describe the canonical model of the production of the learning process as follows:

$$Y_{icgsa} = Y_a\left[\boldsymbol{X}_{icgs}(a), \boldsymbol{S}_s(a), \boldsymbol{C}_{\boldsymbol{cgs}}(a), \theta_{i0}, \varepsilon_{icgsa}\right] \quad (1)$$

where Y_{icsa} is a measure of achievement for child i in classroom c, in grade g, in school s at age a. Acquisition of knowledge is modelled as a combination of cumulative family-supplied inputs ($\boldsymbol{X}_i(a)$), cumulative school-level inputs ($\boldsymbol{S}_s(a)$) such as school management etc., cumulative classroom inputs such as the teacher ($\boldsymbol{C}_{cgs}(a)$) and genetic endowments (θ_{i0}). ε_{icgsa} allows for measurement error in the achievement variable. Y_a allows the impact of all factors to depend on 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; e.g. Rivkin, Hanushek, and Kain 2005, Todd and Wolpin 2003. Treating the arguments in equation (1) as additive separable and assuming that the parameters are not varying with age, equation (1) reduces to:

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

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. A large variance in the effectiveness of teachers suggests a potential for increasing average learning by moving the lowest performing teachers to the level of the best performing teacher. How would a teacher training program such as the NULP affect the variance of classroom/teacher effects? On the one hand, we would expect low performing teachers to have more room for improvement, such that providing teacher training would benefit these teachers relatively more than already high performing teachers. In this case the variance of the teacher effects would decrease as a result of teacher training. On the other hand, we would also expect that high ability teachers would be more able to take advantage of the training provided and thus benefit relatively more than lower ability teachers. In that case the variance of the teacher effects would increase as a result of teacher training. Accordingly, the effect of teacher training on the variance of the teacher effects is an empirical question and depends on the relative strength of the before mentioned effects.

4.2 Empirical Strategy

Classroom Effects

We start our analysis by estimating classroom effects using the following "lagged-score" value-added model[14]:

$$\begin{aligned} Y_{icgst} = \beta_0 + \beta_1 Y_{icgst-1} + \boldsymbol{X'}_{icgst}\beta_2 \\ + \beta_3 \bar{Y}_{-cgst-1} + \lambda_{cgst} + \zeta_g + \tau_t + \zeta_g \tau_t + \beta_4 Y_{icgst-1}\tau_t + \varepsilon_{icst} \end{aligned} \tag{3}$$

where Y_{icgst} is the EGRA testscore for child i in classroom c, in grade g, in school s, in year t. Y_{icst-1} is the EGRA test score from the previous year and captures previous family, school and individual factors as well as genetic endowments (θ_{i0}).[15] $\boldsymbol{X}_{icgst}$ is a vector of individual characteristics and includes gender and age. $\bar{Y}_{-cgst-1}$ is the average baseline score of the classroom peers. λ_{cgst} is the effect of being in a specific classroom and thus $\hat{\lambda}_{cgst}$ is an estimate of the increase in learning attributable to a specific classroom and teacher in year t. We include grade (ζ_g) and year (τ_t) fixed effects as well as allowing the effect of previous test scores and grade-level to vary with time (t).

[14] In a simulation exercise, (Guarino, Reckase, and Wooldridge 2015) find, that the "lagged-score" model performs best in most scenarios. Our results are robust to using a "gain-score" model.

[15] As discussed above, for P1 students we use the baseline scores where available, and otherwise set Y_{icst-1} equal to zero.

To estimate λ_{cgst}, three issues arise: First, there may be school effects that co-vary with true classroom effects, due to factors such as school management, quality that vary at the school-level. Second, there may be individual student effects that co-vary with true classroom effects, due to 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 relatively small samples of students. As the sample gets smaller (fewer students tested per class) the sampling error increases. 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 issues in turn.

(i) Purging school effects from classroom effect estimates

When estimating equation (3) we use both within- and between-school variation. This means that the estimate, $\hat{\lambda}_{cgst}$, picks up both classroom effects and school effects that co-vary with the classroom effects. To overcome this issue we rescale the classroom effects $\hat{\lambda}_{cgst}$ relative to the school mean and thereby only consider the within-school variation in the classroom effects (e.g. Slater, Davies, and Burgess 2012, Araujo et al. 2016, Chetty et al. 2011). This approach nets out all school level factors and thereby provides a lower bound to the degree of variation in the classroom effects.

(ii) Sorting of students to teachers

Endogenous sorting of students to teachers can potentially introduce bias to the value-added approach (see e.g. Rothstein (2010), Kinsler (2012), Chetty, Friedman, and Rockoff (2014) and Goldhaber and Chaplin (2015) for discussions of the severity of this bias). We address this potential source of bias by restricting our sample to the years (2013, 2016 and 2017) when students were randomly assigned to teachers. Two threats to the validity of this approach would be if students systematically switched classrooms during the year, or if student dropout was correlated with teacher ability. We find no evidence of student attrition being systematically related to teacher characteristics (see Appendix Table C.1).

Because we have years in our study when students were randomized to teachers, and years when there was no randomization, we can compare the estimated classroom effects to get a sense of the severity of this bias. To do so, we restrict the sample of teachers to the ones teaching in random assignment years (2013, 2016 and 2017) as well as business-as-usual assignment years (2014 and 2015) and test the difference of the variance of the classroom and teacher effects.

(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 issue we analytically adjust the variance of the estimated classroom effects following the approach suggested by Araujo et al. (2016).[16] For the within-school classroom effects we estimate the variance of the measurement error and subtract that from the estimated variance of the demeaned classroom effects:

$$\hat{V}_{corrected}\left(\hat{\gamma}_{cgst}\right) = V\left(\hat{\gamma}_{cgst}\right) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{\left[\left(\sum_{c=1}^{C_s} N_{cs}\right) - N_{cs}\right]}{N_{cs}\left(\sum_{c=1}^{C_s} N_{cs}\right)}\hat{\sigma}^2\right\} \tag{4}$$

where $\hat{\sigma}^2$ is the variance of the residuals, ε_{icst} from equation (3). *C* is the overall number of classrooms in the sample.

For the estimates of classroom effects that are not demeaned and thus also use between-school variation this expression reduces to:

$$\hat{V}_{corrected}\left(\hat{\gamma}_{cgst}\right) = V\left(\hat{\gamma}_{cgst}\right) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{1}{N_{cs}}\hat{\sigma}^2\right\} \tag{5}$$

Teacher effects

[16] 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.

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 certain assumptions it is possible to separate the teacher effect from classroom effects. The identifying assumption is that any sorting of students to teachers is not systematically occurring year after year. Due to random assignment, this is not a problem in the specifications restricted to 2013, 2016 and 2017. We estimate teacher effects using the demeaned classroom effects with the following equation:

$$\hat{\gamma}_{cgst} = \hat{\alpha}_0 + \hat{\delta}_{cgs} + \omega_{cgst} \qquad (6)$$

where, $\hat{\delta}_{cgs}$ is a vector of teacher indicators and can be interpreted as the "permanent" teacher component. $\hat{\delta}_{cgs}$ are our coefficients of interest when discussing the teacher effects. One important aspect to keep in mind is that with this approach we assume that all time variation in the classroom effects is due to transitory classroom shocks and not "real" teacher quality fluctuations. If this assumption fails ω_{cgst} could contain "real" teacher quality fluctuations and thus bias our teacher effects ($\hat{\delta}_{cgs}$). To address the severity of this bias we also estimate equation (6) using only two adjacent years as this would limit the potential for changes in teacher quality This assumption is more likely to be violated when considering a longer time period where teachers improve over time. As teachers are nested within schools in our sample, sampling error is likely to be correlated over time as the same teachers are teaching in the smaller schools with smaller class sizes. This means that we need to correct the variance of the teacher effects for sampling variation, which we do in the same manner as described above for the classroom effects.

Value-Added Correlations with Teacher Characteristics and Behaviors

To understand the characteristics and behaviors of the most effective teachers, we examine the correlations with our estimated value-added measures. First, we examine if teacher characteristics can explain variation in our estimated measure of teacher effectiveness. We estimate the following equation:

$$\hat{\delta}_{cs} = \beta_0 + \boldsymbol{C}'_{cgs}\beta_1 + \psi_{cgs} \tag{7}$$

where $\hat{\delta}_{cgs}$ are our estimated teacher effects from equation (6), $\boldsymbol{C}_{cgs}$ is a vector of teacher characteristics and includes; gender, years of experience, monthly salary and years of schooling.

Second, we examine if our estimated measure of teacher effectiveness correlates with teacher behavior in the classroom. We use the classroom observations to relate teacher effectiveness to different aspects of teacher behavior including, classroom management and teaching practices as well as student participation. We analyze the data at the level of a 10-minute observation block. Our regression model is:

$$B_{blrcs} = \beta_0 + \beta_1\hat{\gamma}_{cgs} + \boldsymbol{C}'_{cgs}\beta_2 + \rho_s + \varsigma_r + \varphi_{rcs} + \omega_{blrcgs} + \mu_{lrcs} + \epsilon_{blrcs} \tag{8}$$

where *s* indexes schools, *c* indexes classrooms, *r* indexes the round of the visit, *l* indexes the lesson being observed[17], and b indexes the observation block (ie. 1, 2 or 3). Our dependent variables include time use, measures of classroom management constructed through factor analysis, as well as elements of focus, student participation, and materials , B_{blrcs}. Data on teacher behaviors is only available in 2013 and thus our sample of teachers is reduced. To avoid further reduction in our sample by requiring teachers to have multiple years of data we use the estimated classroom effects ($\hat{\gamma}_{cgs}$) as our measure of teacher effectiveness instead of the teacher effects. Moreover, we also include: school (ρ_s), observation round (ς_r) (i.e. indicators of an observation occurring in July, August or October), enumerator (φ_{rcs}), observation block (ω_{blrcgs}) and day-of-the-week (μ_{lrcs}) fixed effects. ϵ_{blrcs} is a mean-zero error term. We cluster the standard errors at the school-level and weight by the share of time spend on reading during the observation window. β_1 is our coefficient of interest and measures how classroom actions vary with teacher effectiveness.

[17] These lessons include: reading (37% of lessons), writing(39% of lessons), English (22% of lessons), Math (2% of lessons) and Other (1% of lessons).

5. Results: Estimates of Teacher Effectiveness

5.1 Full and Longitudinal Samples

Columns 1 and 2 of Table 3 presents the estimates from equations (2) and (7) among all schools available (pooled sample) and columns 3 and 4 present the estimates from the same equations using only schools in the control study arm (*control sample*). This distinction provides information on teacher effectiveness in two different settings: One where school and teacher interventions are common (*pooled sample*) and one where there are no school or teacher interventions (control sample).

Columns 1 and 3 present classroom value-added which is calculated using all teachers available (*full sample*) from equation (2) whereas columns 2 and 4 present teacher value-added which is calculated using teachers with at least two years of data (*longitudinal sample*) using equation (7). We summarize each of the estimates of classroom and teacher value-added measures in terms of standard deviations of student performance on endline exams. We present all estimates with and without corrections for sampling variance. Moreover, we present cluster bootstrapped confidence intervals in square brackets.

[Table 3 about here]

Panel A shows results using both between- and within-school variation to estimate classroom and teacher effects. We find a substantial amount of variation across classrooms and teachers. A one SD increase in teacher quality increases student performance by 0.36-0.52 SDs for the *pooled sample* (columns 1 and 2) and by 0.24-0.41 SDs for the *control sample* (columns 3 and 4). However, because these estimates also include between school variation, some proportion of the variation is likely to be due to non-random sorting of teachers to schools. By implication, these estimates are upper bounds on the variance of true γ_{cgst} (classroom effects) and δ_{cgs} (teacher effects).

To purge the variation of school-level effects, in Panel B we limit the variation to only within-school, effectively comparing teachers between classes in the same year and school. Using this specification we still find substantial variation between teachers, although with smaller magnitudes. The most restrictive result for the *pooled sample* in Column 2 shows that a one SD increase in teacher quality is associated with an increase in student performance by 0.19 SDs and by 0.18 SDs in the *control sample*.

5.2 Random Assignment of Students to Classrooms

To address the potential bias stemming from non-random assignment of students to teachers, we restrict our sample for Table 4 to the years where students were randomly assigned to teachers: 2013, 2016 and 2017. First looking at the estimates purged of school effects in Table 4 (Panel B), we find that a one SD increase in classroom effectiveness increases student performance by 0.29 SDs in the *pooled sample* (Column 1) and by 0.24 SDs in the *control sample* (Column 3). Moving to the teacher value-added we find that a one SD increase in teacher effectiveness increases student performance by 0.14 SDs in the *pooled sample* (Column 2) and by 0.07 SDs in the *control sample* (Column 4). Overall, we see that restricting the sample to random assignment years reduces the variance of value-added estimates.

[Table 4 about here]

In section 6, we directly test how a comprehensive teacher training and pedagogy program affect the variation in value-added estimates, by estimating the impact of the NULP on the variance of the value-added estimates.

5.3 How Biased are Value-added Estimates under Business-as-usual Assignment of Students to Classrooms?

To investigate the degree of bias due to sorting of students to classes we first restrict our full sample to teachers present in both business-as-usual assignment years as well as random assignment years (N=288). Then we split the sample into business-as-usual assignment and random assignment and estimate both classroom and teacher value-added and present the results in Table 5. Columns 1 and 2 present the classroom effects under business-as-usual and random assignment, respectively. Comparing the results in these two columns we see that the random assignment estimates are substantially larger than the those under business-as-usual, consistent with that higher quality teachers being matched with lower performing students.

Because teacher effects are estimated as the teacher-level average of classroom effects across years, if sorting does not systematically occur each year, teacher effects will be less prone to bias based on non-random student sorting as this bias would be purged as a transitory year effect. Indeed the difference between the random assignment and business-as-usual estimates is smaller when comparing the standard deviation of the teacher effects in Columns 3 and 4 (Table 5). However, when formally testing the difference in the value-added estimates between random assignment and business as usual years we reject that the estimates are the same for both the classroom and teacher effects. Thus, in the subsequent sections we will only use the random assignment sample.

5.4 Robustness

In this section we address three issues: a) The imputation of grade-one baseline scores in 2015 and 2016, b) compliance with random assignment in 2013, 2016 and 2017 and c) estimation error from small class sizes.

As mentioned in Section 3.3, baseline scores were not collected in 2015 and 2016 which led us to impute all grade-one baseline scores in 2015 and 2016 with the median grade-one score (which is zero) in 2013 and 2014. While imputing the baseline scores for grade one in 2015 and 2016 allows us to retain a larger sample of teachers over time it also by implication adds non-classical measurement error to our outcome variable and thus potentially biases our estimates. To address the sensitivity of our results, we present two robustness checks in Table 6. First, we omit grade-one students in 2015 and 2016 and re-estimate our main results using the *random assignment samples* – essentially re-running the estimates for Table 4. Second, we replace grade-one baseline scores in all years with zero – including students in 2013 and 2014 for whom we have baseline test scores – and re-estimate the results again using the *random assignment samples*.

[Table 6 about here]

Columns 1 and 2 in Table 6 show that excluding all imputed grade-one scores decreases the standard deviation of the within-school teacher value-added slightly to 0.11 SDs compared to 0.13 SDs in Table 4. Columns 3 and 4 in Table 6 show that replacing all grade-one baseline test scores with zero barely changes the results compared to Table 4. Thus, the decrease in columns 1 and 2 in Table 6 is more likely due to the change in sample than the imputation of grade-one baseline scores. We are therefore not concerned that the imputation of grade-one baseline scores drives our results.

To assess the degree of non-compliance with the random assignment of students to classes in 2013, 2016 and 2017 we test the difference in baseline test scores between streams. We can reject baseline balance in 3.7% of cases, which is below the expected fraction of 5%. Still, we assess the sensitivity of our results in Table 7 and re-estimate the results from Table 4, omitting the school-year-grades for which we can reject baseline balance.

[Table 7 about here]

Table 7 yields similar results as in Table 4 and shows no significant differences compared to the results in Table 4, mitigating some of the concern that our results are sensitive to non-compliance with random assignment for students to classrooms.

As mentioned, our dataset consist of a (random) sample of students within each classroom. This means that we in some cases have a rather small number of students per teacher and as the consistency of the value-added estimates depend on the number of students per teacher this could potentially affect our results[18]. To assess the sensitivity of the inclusion of small class sizes on our results we re-estimate our results from Table 4, omitting class sizes below 10 and 25 students per teacher. Table 8 presents the results.

[Table 8 about here]

Table 8 shows that excluding classrooms with less than 10 students per teacher only have a small effect on the variance of the classroom and teacher effects (columns 1 and 2). However, when excluding classrooms with less than 25 students per teacher (columns 3 and 4) we do see a decrease in the variance of the classroom effects as well as a (smaller) increase in the variance of the teacher effects. This suggests that our results are robust to the exclusion of very small class sizes, but that the results are sensitive to reducing the sample to only including classrooms with at least 25 students per teacher.

[18] 14% of the classrooms have less that 10 students per teacher and 60% have less than 25 students per teacher.

5.5 Correlation with Teacher Characteristics and Behaviours

Using data from the teacher surveys (available in 2013, 2014, 2015 and 2017) and classroom observations (available in 2013), we describe how teacher characteristics and behaviors correlate with higher value-added measures. First, we find that classroom value-added is positively correlated with years of schooling, yet negatively correlated with salary (Table 9). We find no correlation with experience or gender.

[Table 9 about here]

Next, we examine how classroom observation data correlate with teacher value-added, equation (12) in Tables 10 through 12. Table 10 shows the relationship between teacher effectiveness and pedagogical practices in lessons where the students do any reading. Panel A presents the results from estimating the relationship between classroom value-added and the elements of focus in the lesson as well as the degree of participation of the students. We find that more-effective teachers spend less time on letters and words compared to less effective teachers. Panel B presents the results from estimating the relationship between classroom value-added and teaching methods and materials used. Here we find no statistically significant relationship between teacher effectiveness and materials used.

[Table 10 about here]

Table 11 considers the relationship between classroom value-added and pedagogical practices in lessons where the students do any writing. Table 11 is structured the same way as Table 10. In panel A, we find no statistically significant relationship between classroom value-added and the focus on writing elements. In panel B we find that more effective teachers are associated with students spending more time on "air writing"[19], but less time on practicing handwriting. In addition, we find that more effective teachers have students using slates more.

[Table 11 about here]

Table 12 shows the association between teacher effectiveness and speaking/listening behaviors of the students. We find no statistically significant relationship between classroom value-added and student-to-student or student-to-teacher interactions.

[Table 12 about here]

In sum, we find limited correlation between classroom value-added and teacher or student behaviors in the classroom.

6. Effects of the NULP

6.1 Classroom and Teacher Value-added

So far, our analysis has followed the value-added literature by providing estimates of classroom and teacher value-added in an African context. In this section we take the literature further by estimating the impact of a randomized intervention of a comprehensive teacher training and pedagogy program on the variation in teacher effectiveness. While previous literature is able to estimate the scope for test score improvements by (hypothetically) moving the worst performing teachers to the level of the best, we are able to show what actually happens to the value-added estimates when we move teachers through comprehensive training and support.

In Tables 13 and 14, we show how the variance of our classroom and teacher value-added estimates is affected by the introduction of the NULP. Table 13 presents the classroom value-added estimates using the *random assignment sample*.

[Table 13 about here]

[19] Air writing means tracing out the shapes of the letters in the air.

Column 1 shows the results for the group of schools that did not get the program, which is equivalent to Column 3 in Table 4. Columns 2 and 3 present the results from the reduced-cost program and the full-cost program, respectively. The results in Table 13 reveal that the program increases the variance of the classroom effects. Table 14 presents the teacher value-added estimates using the *longitudinal random sample*.

[Table 14 about here]

Table 14 can be interpreted in the same manner as Table 13 and confirms the results that the full-cost program increases the variance of teacher effectiveness. This finding that a highly effective teacher training program is increasing the spread of teacher effectiveness means that some teachers improve more than others. Since the program leads to gains in student performance on average, the most intuitive explanation is that the impact of the program was largest for the highest-quality teachers. A very strict version of this interpretation requires rank preservation. Meaning that, for example, a teacher that belongs to the median for some outcome distribution in the full-cost program, should also have as her counterfactual the median outcome in the control group distribution. To test an implication assumption we follow Djebbari and Smith 2008 and Bitler, Gelbach, and Hoynes 2008 and test whether fixed covariates have same means in a given quantile of the teacher value-added distribution. Table 15 presents the results of that test.

[Table 15 about here]

Each column represent a fixed teacher background variable (including age,gender, salary, experience and years of schooling). Each row correspond to one quartile of the above mentioned outcome distributions. For each quartile of each variable we test the null of zero difference in population quartile means between the full-cost program and the control group (corresponding to 4x7=28 tests). Under the (incorrect) assumption of independence of the different tests, we would expect about two or three rejections. We obtain zero rejections, when using the teacher value-added estimates and two rejections (7%) when using the classroom value-added estimates thus below what we would expect at the 10% level. This provides suggestive evidence for consistency with the rank preservation assumption. One caveat is that we do not have characteristics on all our teachers, so we cannot test this using the full sample of teachers.

6.2 Correlation with Teacher Characteristics

We now investigate how (if at all) the relationship between teacher effectiveness and teacher characteristics differs between treatment arms. One could imagine that providing training and support to teachers could either increase or decrease the correlation of observed characteristics with teacher effectiveness. One the one hand, it could be that having more experience or years of schooling would enable teachers to better take advantage of the training and support provided by the NULP. On the other hand, it could be that the NULP would make characteristics such as experience or education level less important for being an effective teacher. Table 16 presents the results from estimating the effect of the NULP on the relationship between teacher characteristics and classroom value-added by interacting teacher characteristics with indicators for teaching in a reduced-cost or full-cost program school and Table 17 presents the relationship between teacher characteristics and teacher value-added.

[Table 16 about here]

[Table 17 about here]

The results in Table 16 show no differential effect of the NULP on the relationship between classroom value-added and teacher characteristics. Table 17 reveals that the statistically insignificant result between teacher value-added and years of education found in Table 9 masks a statistically significant positive relationship for the teachers in the full-cost program.

7. Conclusion

We use data from a randomized evaluation of a program delivering teacher training and support in northern Uganda to assess the variation in the effectiveness of teachers. The data allows us to make three important contributions to the understanding of teacher effectiveness in low income countries. First, this paper provides the first estimates of teacher effectiveness using the value-added approach in an African country. Utilizing the fact that students were randomly assigned to teachers we can overcome typical issues with bias due to sorting of students to teachers. Second, we are among the first in a developing country able to shed some light on what effective teachers actually do in the classroom. Third, we are able to shed light on how a high impact teacher training program affects the spread of the teacher quality distribution.

Despite severe problems with teaching quality we found that teachers do matter for student learning in northern Uganda. In particular we found that a one standard deviation increase in teacher effectiveness increase student performance by 0.13 to 0.29 standard deviations using a sample of students randomly assigned to teachers and correcting for sampling error. Our upper bound estimate takes both within-school as well as between-school variation into account while our lower bound estimate only considers within-school between-teacher variation. Our lower bound estimate of teacher effectiveness of 0.13 standard deviations is very similar that found for primary schools in the US 0.08 standard deviations Chetty, Friedman, and Rockoff (2014) and Ecuador 0.09 standard deviations Araujo et al. (2016), and slightly lower to that found in Pakistan 0.16 standard deviations Bau and Das (2017). This suggests that teachers are at least as important in a low income context such as Uganda as they are in both high and middle income contexts.

In order to transform the knowledge that "teachers matter" into knowledge that would be useful for policymakers 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 that more years of education are associated with higher teacher effectiveness, while a higher salary is associated with lower teacher effectiveness. We find limited associations between teacher effectiveness and teacher or student behaviors in the classroom. Teacher training and support as provided by the NULP increased test scores on average, but it also increased the spread of the teacher quality distribution making teachers more diverse in their effect on affect student learning. This result that teacher training and support have an outsized impact on the most-effective teachers suggests that an important avenue for future research is to look at how to better reach the less-effective teachers.

References

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *The Quarterly Journal of Economics* no. 131:1415-1453. doi: 10.1093/qje/qjw016.

Azam, Mehtabul, and Geeta Gandhi Kingdon. 2015. "Assessing teacher quality in India." *Journal of Development Economics* no. 117:74-83. doi: 10.1016/j.jdeveco.2015.07.001.

Bau, Natalie, and Jishnu Das. 2017. "The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers." *World Bank Policy Research Working Paper*.

Bilker, Warren B., John A. Hansen, Colleen M. Brensinger, Jan Richard, Raquel E. Gur, and Ruben C. Gur. 2012. "Development of abbreviated nine-item forms of the Raven's standard progressive matrices test." *Assessment* no. 19:354-369. doi: 10.1177/1073191112446655.

Black, Dan A., and Jeffrey A. Smith. 2006. "Estimating the Returns to College Quality with Multiple Proxies for Quality." *Journal of Labor Economics* no. 24:701-728. doi: 10.1086/505067.

Bold, Tessa, Deon P. Filmer, Gayle Martin, Molina Ezequiel, Christophe Rockmore, Brian William Stacy, Kristina Svensson, and Waly Wane. 2017. "What do teachers know and do ? does it matter ? evidence from primary schools in Africa." *Policy Research working paper*.

Buhl-Wiggers, Julie, Jason T. Kerwin, Jeffrey Smith, and Rebecca Thornton. 2018. *Program Scale-up and Sustainability*. (Working Paper).

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics* no. 126:1593-1660. doi: 10.1093/qje/qjr041.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* no. 104:2593-2632. doi: 10.1257/aer.104.9.2593.

Lee Crawfurd. 2017. School Management and Public–Private Partnerships in Uganda

Deininger, Klaus. 2003. "Does cost of schooling affect enrollment by the poor? Universal primary education in Uganda." *Economics of Education Review* no. 22:291-305. doi: 10.1016/S0272-7757(02)00053-5.

Djebbari, Habiba, and Jeffrey Smith (2008). "Heterogeneous impacts in PROGRESA" *Journal of Econometrics*. no 145. pp 64-80

Dubeck, Margaret M., and Amber Gove. 2015. "The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations." *International Journal of Educational Development* no. 40:315-322. doi: 10.1016/j.ijedudev.2014.11.004.

Evans, David K., and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *The World Bank Research Observer* no. 31:242-270. doi: 10.1093/wbro/lkw004.

Ganimian, Alejandro J., and Richard J. Murnane. 2014. Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Impact Evaluations. National Bureau of Economic Research.

Glewwe, P., and K. Muralidharan. 2016. "Improving Education Outcomes in Developing Countries." *Handbook of the Economics of Education* no. 5:653-743. doi: 10.1016/B978-0-444-63459-7.00010-5.

Glewwe, Paul, Phillip H. Ross, and Bruce Wydick. 2017. "Developing Hope Among Impoverished Children: Using Child Self-Portraits to Measure Poverty Program Impacts." *Journal of Human Resources*:0816-8112R1. doi: 10.3368/jhr.53.2.0816-8112R1.

Goldhaber, Dan, and Duncan Dunbar Chaplin. 2015. "Assessing the “Rothstein Falsification Test”: Does It Really Show Teacher Value-Added Models Are Biased?" *Journal of Research on Educational Effectiveness* no. 8:8-34. doi: 10.1080/19345747.2014.978059.

Gove, Amber, and Anna Wetterberg. 2011. *The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy*: RTI International.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review* no. 100:267-271. doi: 10.1257/aer.100.2.267.

Hardman, Frank, Jim Ackers, Niki Abrishamian, and Margo O'Sullivan. 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* no. 41:669-683. doi: 10.1080/03057925.2011.581014.

Horvath, Hedvig. 2015. "Classroom Assignment Policies and Implications for Teacher Value-Added Estimation." *Unpublished Manuscript*.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. National Bureau of Economic Research.

Kerwin, Jason, T., and Rebecca Thornton. 2017. "Making the Grade: The Trade-off between Efficiency and Effectiveness in Improving Student Learning."

Kim, Thomas, and Saul Axelrod. 2005. "Direct instruction: An educators' guide and a plea for action." *The Behavior Analyst Today* no. 6:111-120. doi: 10.1037/h0100061.

Kinsler, Josh. 2012. "Assessing Rothstein's critique of teacher value-added models." *Quantitative Economics* no. 3:333-362. doi: 10.3982/QE132.

Cory Koedel and Julian R. Betts Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique Education Finance and Policy 2011 6:1, 18-42

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The challenge of education and learning in the developing world." *Science (New York, N.Y.)* no. 340:297-300. doi: 10.1126/science.1235350.

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* no. 85:353-394. doi: 10.3102/0034654314553127.

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

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* no. 73:417-458. doi: 10.1111/j.1468-0262.2005.00584.x.

Rothstein, Jesse. 2009. Student sorting and bias in value-added estimation: Selection on observables and unobservables. Education Finance and Policy 4(4): 538–72.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *The Quarterly Journal of Economics* no. 125:175-214. doi: 10.1162/qjec.2010.125.1.175.

RTI. 2009. Early Grade Reading Assessment Toolkit. World Bank Office of Human Development.

Slater, Helen, Neil M. Davies, and Simon Burgess. 2012. "Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England*." *Oxford Bulletin of Economics and Statistics* no. 74:629-645. doi: 10.1111/j.1468-0084.2011.00666.x.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* no. 113:F3-F33. doi: 10.1111/1468-0297.00097.

Ugandan Ministry of Education and Sports. 2014. Teacher Issues in Uganda: A shared vision for an effective teachers policy. UNESCO - IIEP Pôle de Dakar.

Uwezo. 2016. Are Our Children Learning (2016)? Uwezo Uganda Sixth Learning Assessment Report. Kampala: Twaweza East Africa.

World Bank. *World Development Indicators 2010*, 2010 2010.

World Bank. 2013. "World Developemnt Indicators 2013."

Figures and Tables

Table 1: School, Teacher and Pupil Samples

	2013	2014	2015	2016	2017
Teacher Samples					
Full Sample					
#Schools	37	64	128	128	128
#Teachers	79	164	436	563	430
#Pupils	1,430	2,904	10,801	14,993	9,783
Pupils/Teacher	20	22	32	37	29
Longitudinal Sample					
#Schools	32	63	125	128	105
#Teachers	54	118	302	389	171
#Pupils	1,029	2,290	7,636	11,019	4,110
Pupils/Teacher	20	23	32	38	30
Random assignment of studnets	Yes	No	No	Yes	Yes
Grades assesed	P1	P1, P2	P1 - P3	P1 - P4	P3 - P5

Notes: The Full Sample includes all teachers available in schools where there are at least two teachers. The Longitudinal Sample includes all teachers who are teaching in at least two different years.

Table 2: Descriptive Statistics

		Full Sample			Longitudinal Sample	
Students	*Control*	*Reduced-cost*	*Full-cost*	*Control*	*Reduced-cost*	*Full-cost*
Age	9.05	9.07	9.02	8.73	8.77	8.66
Female (%)	49.52	50.02	49.68	48.37	51.33	49.81
Baseline score	0.00	0.36	0.65	0.00	0.31	0.56
Endline score	0.05	0.60	1.05	0.03	0.57	1.01
N	12,103	13,908	13,900	7,469	9,221	9,394
Teachers						
Age	39.52	41.10	39.72	39.57	42.06	39.56
Female (%)	50.90	44.25	38.65	56.79	44.14	41.58
Salary (shillings)	442,357	427,345	431,844	416,299	420,298	403,651
Yrs of experience	14.10	15.06	14.31	14.37	15.72	14.19
Yrs of education	14.63	14.48	14.56	14.55	14.48	14.53
Ravens score	1.91	1.86	1.99	1.87	1.89	2.01
N	354	380	351	135	159	153

		Randomized Teachers Sample			Longitudinal Randomized Teachers Sample	
Students	*Control*	*Reduced-cost*	*Full-cost*	*Control*	*Reduced-cost*	*Full-cost*
Age	9.34	9.43	9.40	9.51	9.71	9.56
Female (%)	49.38	50.17	49.75	48.22	51.58	50.14
Baseline score	0.00	0.40	0.71	0.00	0.44	0.89
Endline score	0.05	0.61	1.07	0.07	0.69	1.36
N	7,890	9,014	8,549	2,748	3,232	2,900
Teachers						
Age	39.77	40.62	39.50	40.07	39.91	39.17
Female (%)	47.46	41.41	37.34	55.83	42.67	40.76
Salary (shillings)	463,548	439,080	451,221	445,692	445,894	431,580
Yrs of experience	14.29	14.76	13.75	15.18	14.84	12.35
Yrs of education	14.63	14.46	14.53	14.52	14.56	14.54
Ravens score	1.91	1.86	2.01	2.26	2.09	2.35
N	275	304	281	58	67	60

Notes: The full sample includes all teachers available. The longitudinal sample includes all teachers who are teaching in at least two different years (from 2013-2017). The randomized teachers sample includes all teachers teaching in either 2013, 2016 or 2017 when students were randomly assigned to classrooms. The longitudinal randomized teachers sample includes teachers teaching in at least two of the random assignment years (2013, 2016 and 2017).

Table 3: Classroom and Teacher Value-Added Estimates Full Sample and Longitudinal Sample

	All Schools		Control Schools	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Including school effects	(1)	(2)	(3)	(4)
SD of effects	0.56	0.42	0.44	0.30
	[0.39-0.76]	[0.25-0.65]	[0.20-0.66]	[0.02-0.55]
Corrected SD of effects	0.52	0.36	0.41	0.24
	[0.34-0.72]	[0.19-0.61]	[0.15-0.63]	[0.00-0.51]
Panel B: School effects purged				
SD of effects	0.37	0.26	0.34	0.23
	[0.21-0.53]	[0.13-0.41]	[0.16-0.49]	[0.08-0.35]
Corrected SD of effects	0.33	0.19	0.30	0.18
	[0.15-0.49]	[0.04-0.35]	[0.12-0.46]	[0.01-0.32]
Children	39,911	26,084	12,103	7,469
Teachers	1,672	447	525	135
Schools	128	128	42	42
Pupils per classroom/teacher	24	25	23	25
Sample	Full	Longitudinal	Full	Longitudinal

Notes: The Full Sample includes all teachers available in the study schools while the Longitudinal Sample includes teachers available in at least two different years between 2013 and 2017. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Panel B shows estimates purged of school effects by subtracting off the year-specific school mean. Control schools (N=42) did not receive the NULP intervention.

Table 4: Classroom and Teacher Value-Added Estimates Randomized Teachers Sample

	All Schools		Control Schools	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Including school effects	(1)	(2)	(3)	(4)
SD of effects	0.47	0.35	0.34	0.23
	[0.23-0.75]	[0.20-0.58]	[0.00-0.64]	[0.00-0.49]
Corrected SD of effects	0.42	0.29	0.29	0.15
	[0.18-0.72]	[0.13-0.53]	[0.00-0.61]	[0.00-0.44]
Panel B: School effects purged				
SD of effects	0.33	0.22	0.28	0.16
	[0.10-0.55]	[0.05-0.38]	[0.03-0.49]	[0.00-0.32]
Corrected SD of effects	0.29	0.13	0.24	0.07
	[0.04-0.52]	[0.00-0.32]	[0.00-0.46]	[0.00-0.25]
Children	26,206	9,185	8,101	2,842
Teachers	1,072	185	340	58
Schools	128	105	42	34
Pupils per classroom/teacher	24	25	24	25
Sample	Random	Longitudinal Random	Random	Longitudinal Random

Notes: The Random Assignment Sample includes all teachers teaching in 2013, 2016 or 2017 when students were randomly assigned to classrooms. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Panel B shows estimates purged of school effects by subtracting off the year-specific school mean. Control schools (N=42) did not receive the NULP intervention.

Table 5: Comparison of Random Assignment and Business-as-usual Value-Added Estimates

	Classroom Effects			Teacher Effects	
	Business-as-usual	Random assignment		Business-as-usual	Random assignment
Panel A: Including school effects	(1)	(2)		(3)	(4)
SD of effects	0.62	0.49		0.55	0.44
	[0.18-0.24]	[0.27-0.37]		[0.19-0.23]	[0.18-0.24]
Corrected SD of effects	0.57	0.44		0.5	0.39
	[0.15-0.21]	[0.24-0.34]		[0.16-0.21]	[0.13-0.20]
Panel B: School effects purged					
SD of effects	0.41	0.34		0.3	0.23
	[0.13-0.18]	[0.23-0.32]		[0.14-0.18]	[0.14-0.18]
Corrected SD of effects	0.36	0.28		0.23	0.19
	[0.11-0.15]	[0.20-0.30]		[0.12-0.16]	[0.09-0.14]
Children	8582	9524	0	938	1081
Teachers	288	288	0	21	21
Schools	124	124	0	17	17
Pupils per classroom/teacher	24	27	0	22	25

Notes: The Business-as-usual assignment sample is includes data from 2014 and 2015. The Random assignment sample includes data from 2013, 2016 and 2017. The table only includes teachers that teach in both business-as-usual and random assignment years. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Panel B shows estimates purged of school effects by subtracting off the year-specific school mean.

Table 6: Robustness: Imputation of Grade-One Baseline Scores

	Omitting P1 in 2016 and 2015		Replacing all BL P1 scores with zero	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Including school effects	(1)	(2)	(3)	(4)
SD of effects	0.49	0.45	0.47	0.35
	[0.26-0.71]	[026-0.64]	[0.20-0.73]	[0.12-0.58]
Corrected SD of effects	0.44	0.4	0.42	0.29
	[0.20-0.67]	[0.20-0.60]	[0.15-0.70]	[0.05-0.53]
Panel B: School effects purged				
SD of effects	0.33	0.21	0.33	0.22
	[0.16-0.50]	[0.10-0.32]	[0.12-0.55]	[0.06-0.38]
Corrected SD of effects	0.28	0.11	0.29	0.14
	[0.10-0.46]	[0.00-0.23]	[0.06-0.51]	[0.00-0.31]
Children	19,357	8,173	26,206	9,185
Teachers	747	185	885	185
Schools	128	105	128	105
Pupils per classroom/teacher	21	23	24	25
Sample	Random	Longitudinal random	Random	Longitudinal random

Notes: The Random Assignment Sample includes all teachers teaching in 2013, 2016 or 2017 when students were randomly assigned to classrooms. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Panel B shows estimates purged of school effects by subtracting off the year-specific school mean.

Table 7: Robustness: Compliance with Random Assignment

	All Schools		Control Schools	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Including school effects	(1)	(2)	(3)	(4)
SD of effects	0.48	0.36	0.34	0.24
	[0.23-0.75]	[0.20-0.58]	[0.00-0.64]	[0.00-0.49]
Corrected SD of effects	0.43	0.3	0.29	0.17
	[0.18-0.72]	[0.13-0.53]	[0.00-0.61]	[0.00-0.44]
Panel B: School effects purged				
SD of effects	0.34	0.22	0.28	0.17
	[0.10-0.55]	[0.05-0.38]	[0.03-0.49]	[0.00-0.32]
Corrected SD of effects	0.29	0.15	0.24	0.09
	[0.04-0.52]	[0.00-0.32]	[0.00-0.46]	[0.00-0.25]
Children	25408	8835	7890	2748
Teachers	1031	185	328	58
Schools	128	105	42	34
Pupils per classroom/teacher	25	25	24	25
Sample	Random + baseline balance	Longitudinal random + baseline balance	Random + baseline balance	Longitudinal random + baseline balance

Notes: The Full Sample includes all teachers available in the study schools while the Longitudinal Sample includes teachers available in at least two different years between 2013 and 2016. We include data collected in years where pupils were randomly assigned to classes (2013, 2016 and 2017) and where we cannot reject baseline balance of tests cores. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Panel B shows estimates purged of school effects by subtracting off the year-specific school mean.

Table 8: Robustness: Restricting to classes with minimum 10 or 25 students

	Class size >=10		Class size >=25	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Including school effects	(1)	(2)	(3)	(4)
SD of effects	0.46	0.35	0.38	0.38
	[0.19-0.72]	[0.11-0.59]	[0.12-0.65]	[0.13-0.63]
Corrected SD of effects	0.41	0.29	0.35	0.35
	[0.14-0.69]	[0.01-0.54]	[0.07-0.62]	[0.09-0.61]
Panel B: School effects purged				
SD of effects	0.31	0.22	0.23	0.22
	[0.10-0.53]	[0.06-0.39]	[0.02-0.45]	[0.06-0.38]
Corrected SD of effects	0.27	0.16	0.19	0.18
	[0.04-0.49]	[0.00-0.34]	[0.00-0.42]	[0.01-0.36]
Children	24,987	8,775	14,128	5,021
Teachers	914	178	374	105
Schools	127	103	120	81
Pupils per classroom/teacher	27	27	38	34
Sample	Random	Longitudinal random	Random	Longitudinal random

Notes: The Random Assignment Sample includes all teachers teaching in 2013, 2016 or 2017 when students were randomly assigned to classrooms 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Panel B shows estimates purged of school effects by subtracting off the year-specific school mean.

Table 9: Correlation with Teacher Characteristics

	All years		Random assigment years	
	Teacher Effects	Classroom Effects	Teacher Effects	Classroom Effects
	(1)	(2)	(3)	(4)
Years of Schooling	0.318	0.365**	0.238	0.328**
	(0.199)	(0.167)	(0.195)	(0.163)
Years of Schooling^2	-0.011	-0.013**	-0.008	-0.012**
	(0.007)	(0.006)	(0.007)	(0.006)
Salary (log)	0.137	-0.038	-0.182*	-0.149***
	(0.083)	(0.043)	(0.096)	(0.043)
Male (1=Yes)	0.001	-0.010	0.010	-0.024
	(0.027)	(0.022)	(0.030)	(0.023)
< 5 yrs of experience (1=Yes)	0.055	0.044	0.066	0.045
	(0.069)	(0.044)	(0.078)	(0.040)
Observations	342	665	165	584
R-squared	0.020	0.015	0.034	0.035

Notes: Standard errors are clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001. The dependent variables are teacher and classroom effects.

Table 10: Classroom Observations: Reading Activities

	Element of Focus				
	Sounds	Letters	Words	Sentences	Percent Pupils Participating
	(1)	(2)	(3)	(4)	(5)
Classroom Effects	-0.058	-0.282**	-0.276***	-0.049	2.909
	(0.084)	(0.122)	(0.091)	(0.165)	(2.248)
Observations	521	521	521	521	521
Adjusted R-Squared	.094	.053	.03	.117	.283

	Teaching Method				Materials Used		
	Whole class	Small groups	Individual at seat	Individual at board	Board	Primer	Reader
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Classroom Effects	0.100	-0.099	0.082	0.064	0.074	-0.063	0.062
	(0.111)	(0.078)	(0.208)	(0.142)	(0.151)	(0.086)	(0.087)
Observations	521	521	521	521	521	521	521
Adjusted R-Squared	.068	.159	.052	.077	.152	.192	.195

Notes: Sample is observation windows, based on 223 individual lesson observations for 45 teachers in 30 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for: Teacher gender experience, years of schooling, ravens score and salary as well as indicators for the round of the observations, the period of the observation window (1, 2, or 3), the enumerator, the day of the week, school and are weighted by the share of time spent on reading during the observation window. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

Table 11: Classroom Observations: Writing Activities

	Element of Focus					
	Pictures	Letters	Words	Sentences	Name	Percent Pupils Participating
	(1)	(2)	(3)	(4)	(5)	(6)
Classroom Effects	-0.099	0.065	-0.177	0.196	0.114	3.297
	(0.165)	(0.155)	(0.129)	(0.155)	(0.268)	(4.589)
Observations	309	309	309	309	309	309
Adjusted R-Squared	.047	.125	.194	.168	.2	.183

	Teaching Method				Materials Used		
	Air writing	Handwriting practice	Copy text from board	Writing own text	Board	Slate	Paper
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Classroom Effects	0.222**	-0.277**	0.153	-0.164	0.001	0.383**	-0.141
	(0.089)	(0.123)	(0.220)	(0.139)	(0.077)	(0.153)	(0.134)
Observations	309	309	309	309	309	309	309
Adjusted R-Squared	.089	.336	.26	.166	0.103	.348	.162

Notes: Sample is observation windows, based on 186 individual lesson observations for 45 teachers in 30 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for: Teacher gender, experience, years of shooling, ravens score and salary as well as indicators for the round of the observations, the period of the observation window (1, 2, or 3), the enumerator, the day of the week, school and are weighted by the share of time spent on writing during the observation window. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

Table 12: Classroom Observations: Pupils Speaking and Listening

	To Partner	To Small Group	To Whole Class	To Teacher	Percent Pupils Participating
	(1)	(2)	(3)	(4)	(5)
Classroom Effects	0.041	-0.030	-0.108	-0.061	-3.840
	(0.149)	(0.057)	(0.122)	(0.061)	(3.228)
Observation Windows	737	737	737	737	737
Adjusted R-Squared	.252	.092	.228	.081	.332

Notes: Sample is observation windows, based on 246 individual lesson observations for 45 teachers in 30 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for: Teacher gender, experience, years of shooling, ravens score and salary as well as indicators for the round of the observations, the period of the observation window (1, 2, or 3), the enumerator, the day of the week and school. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

Table 13: Heterogeneity of Classroom Value-Added by NULP Study Arm

	Classroom Effects		
	Control	Reduced-Cost	Full-Cost
Panel A: Including school effects	(1)	(2)	(3)
SD of effects	0.34	0.43	0.55
	[0.01-0.62]	[0.18-0.69]	[0.33-0.79]
Corrected SD of effects	0.29	0.38	0.51
	[0.00-0.59]	[0.12-0.66]	[0.27-0.75]
Panel B: School effects purged			
SD of effects	0.28	0.34	0.38
	[0.03-0.48]	[0.12-0.56]	[0.15-0.58]
Corrected SD of effects	0.24	0.29	0.33
	[0.00-0.45]	[0.05-0.52]	[0.09-0.54]
Children	8,101	9,180	8,925
Teachers	340	379	353
Schools	42	44	42
Pupils per classroom/teacher	24	24	25

Notes: The sample includes all teachers available in random assignment years (2013, 2016 and 2017) as well as pass the test of balance in baseline scores. 95% confidece intervals for the SD of the classroom effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.

Table 14: Heterogeneity of Teacher Value-Added by NULP Study Arm

	Teacher Effects		
	Control	Reduced-Cost	Full-Cost
Panel A: Including school effects	(1)	(2)	(3)
SD of effects	0.23	0.28	0.39
	[0.00-0.48]	[0.03-0.54]	[0.22-0.53]
Corrected SD of effects	0.15	0.19	0.32
	[0.00-0.43]	[0.00-0.48]	[0.13-0.47]
Panel B: School effects purged			
SD of effects	0.16	0.23	0.24
	[0.00-0.32]	[0.04-0.41]	[0.07-0.39]
Corrected SD of effects	0.07	0.15	0.15
	[0.00-0.25]	[0.00-0.35]	[0.00-0.31]
Children	2,842	3,285	3,058
Teachers	58	67	60
Schools	34	37	34
Pupils per classroom/teacher	25	24	25

Notes: The sample includes all teachers available in random assignment years (2013, 2016 and 2017) as well as pass the test of balance in baseline scores. 95% confidece intervals for the SD of the teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.

Table 15: Rank Preservation

Panel A: Teacher Effects

Teacher Characteristics

Differenc between full-cost and control group means in:	Age	Gender	Salary	Experience	Schooling
	(1)	(3)	(4)	(5)	(6)
First quartile of TVA	-3.307	0.154	-0.041	-5.235	-0.452
	(3.523)	(0.180)	(0.049)	(3.740)	(0.455)
Second quartile of TVA	-2.671	0.107	-0.038	-1.033	-0.430
	(3.160)	(0.183)	(0.061)	(3.049)	(0.504)
Third quartile of TVA	1.899	-0.050	0.117	1.168	0.389
	(2.879)	(0.207)	(0.091)	(2.767)	(0.446)
Fourth quartile of TVA	2.510	-0.102	-0.009	2.605	0.434
	(2.990)	(0.179)	(0.039)	(2.617)	(0.486)
Observations	178	179	166	167	178

Panel B: Classroom effects

Teacher Characteristics

Differenc between full-cost and control group means in:	Age	Gender	Salary	Experience	Schooling
	(1)	(3)	(4)	(5)	(6)
First quartile of TVA	-0.081	-0.007	0.051	-0.144	-0.263
	(1.747)	(0.097)	(0.055)	(1.625)	(0.244)
Second quartile of TVA	-2.426	0.037	-0.080	-2.118	-0.061
	(1.727)	(0.080)	(0.078)	(1.675)	(0.251)
Third quartile of TVA	-0.732	-0.051	0.030	-0.463	0.024
	(1.642)	(0.089)	(0.033)	(1.427)	(0.251)
Fourth quartile of TVA	2.358*	0.049	0.022	2.510*	0.203
	(1.411)	(0.080)	(0.031)	(1.464)	(0.249)
Observations	745	771	595	716	768

Notes: Robust standard errors in parentheses, clustered by school. All regressions control for stratification cell fixed-effects. *** p<0.01, ** p<0.05, * p<0.1. TVA = Teacher Value Added (using the random assignment sample).

Table 16: Effects of the NULP on the Relationship between Classroom Value-added and Teacher Characteristics

	Classroom Effects			
	(1)	(2)	(3)	(6)
< 5 Yrs of experience	0.054			0.058
	(0.056)			(0.057)
Reduced-cost * < 5 yrs of experience	-0.029			-0.036
	(0.090)			(0.092)
Full-cost *< 5 yrs of experience	0.031			0.026
	(0.101)			(0.101)
Yrs of education		0.314*		0.300*
		(0.163)		(0.163)
Reduced-cost * Yrs of education		-0.010		-0.010
		(0.018)		(0.018)
Full-cost * Yrs of education		0.009		0.007
		(0.019)		(0.020)
Log salary (shillings)			-0.132*	-0.131*
			(0.074)	(0.075)
Reduced-cost * Log salary			0.106	0.098
			(0.122)	(0.123)
Full-cost * Log salary			-0.050	-0.042
			(0.093)	(0.093)
Observations	594	594	594	594
R-squared	0.039	0.040	0.041	0.042

Notes: All regressions control for: Gender, Years of schooling, Experience and Salary. Standard errors are clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

Table 17: Effects of the NULP on the Relationship between Teacher Value-added and Teacher Characteristics

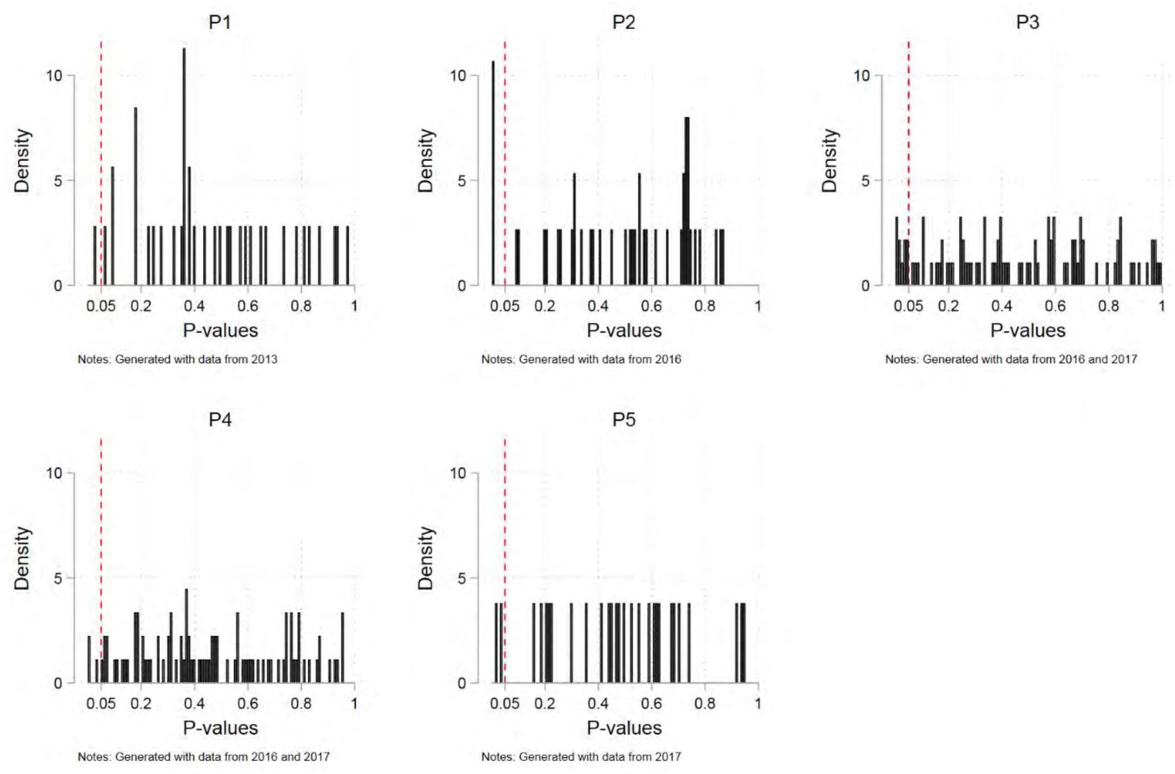
	Teacher Effects			
	(1)	(2)	(3)	(6)
< 5 Yrs of experience	-0.090			-0.079
	(0.061)			(0.063)
Reduced-cost * < 5 yrs of experience	0.187			0.181
	(0.114)			(0.122)
Full-cost *< 5 yrs of experience	0.177			0.177
	(0.162)			(0.160)
Yrs of education		0.073		0.086
		(0.178)		(0.181)
Reduced-cost * Yrs of education		-0.016		-0.017
		(0.020)		(0.021)
Full-cost * Yrs of education		0.070***		0.068***
		(0.025)		(0.025)
Log salary (shillings)			-0.175	-0.191
			(0.130)	(0.130)
Reduced-cost * Log salary			0.063	0.070
			(0.243)	(0.242)
Full-cost * Log salary			-0.011	0.053
			(0.218)	(0.229)
Observations	165	165	165	165
R-squared	0.074	0.106	0.066	0.114

Notes: All regressions control for: Gender, Years of schooling, Experience and Salary. Standard errors are clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

Appendices

Appendix A Verifying Random Assignment

Figure A.1 : Distributions of P-values testing differences in baseline scores between classrooms within each school in random assignment years



Notes: These P-values are calculated from regressing baseline test scores on teacher indicators within each school and testing the difference between teachers using an F-test. When multiple years are pooled the regressions include year fixed effects. The red line marks a P-value of 0.05.

Appendix B Distributions of Baseline Subtests for grade-one in 2013 and 2014

Figure B.1: Distribution of the raw scores in the subtest for grade one in 2013

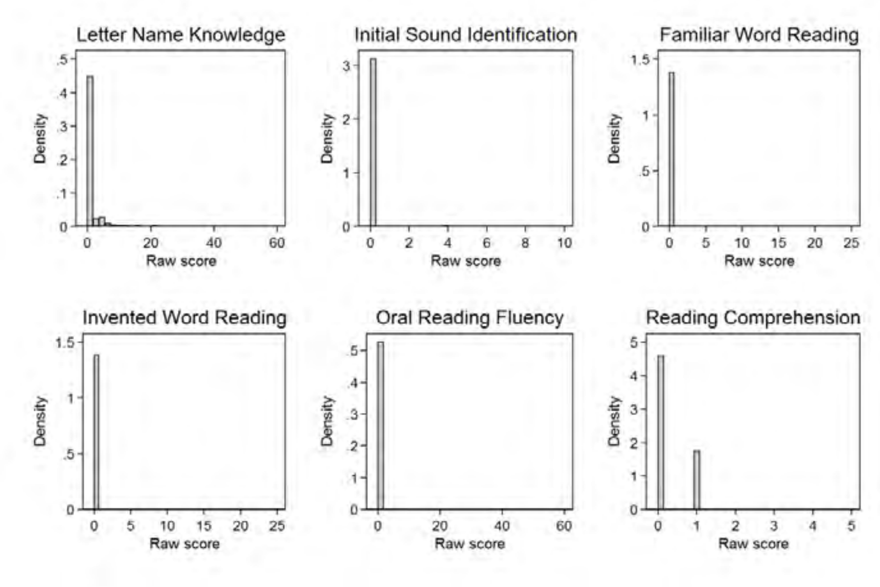
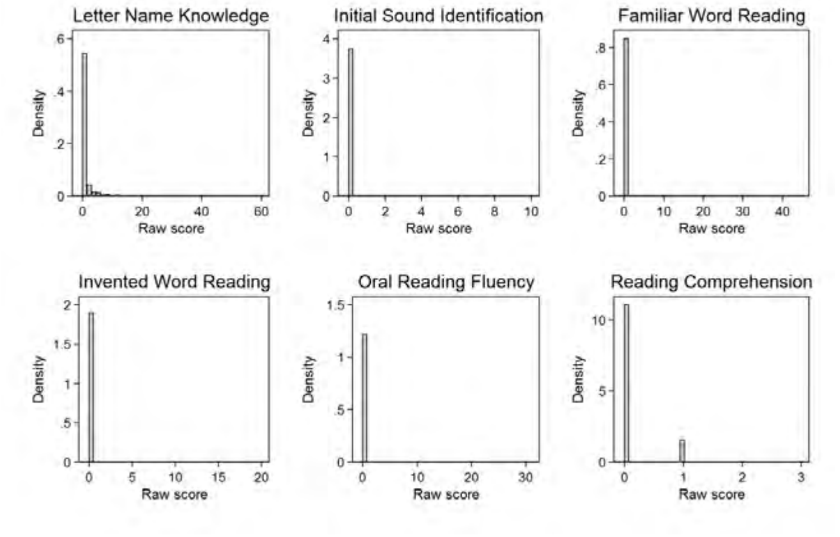


Figure B.2: Distribution of the raw scores in the subtest for grade one in 2014



Appendix C Attrition

Table C.1: Correlation between the Probability of Attritting and Teacher Characteristics

	(1)
Years of schooling	-.005
	(.009)
Observations	19277
Log salary (shillings)	-.078
	(.079)
Observations	19232
Male (yes=1)	-.018
	(.023)
Observations	19480
Experience (years)	.001
	(.001)
Observations	18999

Notes: Dependent variable: Indicator for being an attritor. All regressions control for indicators for year, grade-level and school. Standard errors are clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.