Working paper

International Growth Centre

Teacher effectiveness in Africa

Longitudinal and causal estimates

Julie Buhl-Wiggers
Jason T. Kerwin
Jeffrey A. Smith
Rebecca Thornton

November 2018

When citing this paper, please use the title and the following reference number:
S-89238-UGA-1

DIRECTED BY LSE UNIVERSITY OF OXFORD

FUNDED BY UKaid from the British people

Teacher Effectiveness in Africa: Longitudinal and Causal Estimates

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

November 29, 2018

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 0.09 SD improvement in student performance on a reading test 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 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).

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 (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 (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 developing countries and randomized control trials conducted in developing studies. This paper integrates these two approaches to shed light on the relationship between teachers and student learning in Uganda.

Specifically, the aims of this study are threefold. First, we present the first value-added estimates of teacher effectiveness from an African country; our results are also some of the very 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 a rich set of 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. Which means that we, contrary to previous literature, 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 teacher-level 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. It began 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 cover 128 schools. The evaluation assigned each of the schools 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. An 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).

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 to estimate teacher effectiveness (Chetty et al. 2014, Koedel and Betts 2011, Rothstein 2009). Second, using the randomization of the NULP across schools, we are able to estimate the causal impact of teacher training on the variance of teacher effectiveness. This provides insight to whether teacher training and support make teachers more similar or more varied in their ability to affect student learning.

Our lower-bound estimate of the teacher value added is that a one-standard deviation increase in teacher effectiveness improves test scores by 0.09 standard deviations. These lower-bound estimates are derived from within-school variation, corrected for sampling variation and are strikingly similar to other 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]

As common in the literature, we find no relationship between teacher effectiveness and observed characteristics such as age, experience, or education. However, using a rich set of classroom observations data, we find that more effective teachers are more likely to have a solid lesson plan and to have more active students.

When we evaluate the effects of the NULP intervention, we find a large increase in the spread of classroom value-added. Compared to the control group estimate of 0.03 standard deviations, one standard deviation increase in teacher effectiveness in full-cost program schools leads to an increase in student performance of 0.20 standard deviations.

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, observed teacher characteristics are not sufficient to measure teacher effectiveness and thus screening effective teachers *ex ante* does not seem feasible with traditional measures such education level, experience etc. More research is needed on how to design personal policies based on *ex post* evaluation of teachers or on which alternative characteristics to observe *ex ante*. Third, better teachers appear to gain more from teacher training and support, making it crucial to better understand how to reach the worst performing teachers.

2. Setting and Intervention Details

2.1 Primary Education in Uganda

Primary education in Uganda consists of seven years of education with 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 improvement in access, late

[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. Crawfurd estimates that the "management" value-added from the World Management Survey in Uganda is 0.06 standard deviations (Crawfurd 2017).

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. 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] 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.

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.[4] In 2014 the program was expanded to 90 additional schools for a total of 128 schools. The eligibility criteria for these new 90 schools were slightly different, and less stringent.[5] The number of classrooms per grade was no longer stipulated.

Students and Teachers

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.

3.2 Randomization

Assignment of students to classrooms and teachers

Our research design takes advantage of the fact that students were randomly assigned to teachers. In three years, 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.

[4] 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.

[5] 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.

Compliance with this procedure was verified by having field staff compare the original student lists to the randomized rosters, and also by asking the head teachers what they did. In order to test compliance, we take the approach suggested by Horvath (2015) and compare baseline score means across classrooms within schools and grade level each year. We find that 10% of the classrooms had statistically-significant baseline differences between streams. We do robustness checks excluding those classrooms.[6]

In 2014 and 2015, head teachers were not given explicit instructions on how to assign students or teachers. In general, the way assignments are made is specific to each school, and depends on the approach used by the school's head teacher. In order to assess the degree of sorting present in these years we also test the differences in baseline score means across classrooms within schools and grade level for each year. Here we also find that around 10% of the classrooms had significant baseline differences between streams.

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 Samples

Teachers

We work with three 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

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

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 *Randomized Teacher Sample* is a subsample of the Full Sample, restricted to years where students were randomly assigned to teachers (2013, 2016 and 2017) but teachers are not necessarily teaching in all three years.

Table 1: School, Teacher and Student Samples

	Pooled	*Control*	*Reduced-cost*	*Full-cost*
Full Sample				
# Schools	128	42	44	42
# Teachers	942	301	322	319
# Children	37,649	11,284	13,091	13,274
Pupils/Teacher	34	33	34	34
Logitudinal Sample				
# Schools	127	41	44	42
# Teachers	409	124	143	142
# Children	24,461	6,916	8,768	8,777
Pupils/Teacher	35	34	35	35
Randomized Teachers Sample				
# Schools	128	42	44	42
# Teachers	782	249	266	267
# Children	25,251	7,810	8,762	8,679
Pupils/Teacher	34.48	33.85	35.15	34.38
Longitudinal Randomized Teacher Sample				
# Schools	103	33	36	34
# Teachers	158	50	54	54
# Children	8,339	2,594	2,934	2,811
Pupils/Teacher	33	33	33	32

Notes: The Full Sample includes all teachers available in 128 study schools. The Longitudinal Sample includes all teachers who are teaching in at least two different years (from 2013-2016). The Randomized Teacher Sample includes all teachers teaching in 2013, 2016 or 2017 when students were randomly assigned to classrooms. The Longitudinal Randomized Teacher Sample includes teachers teaching in at least two of the random assignment years (2013, 2016 and 2016).

Our sample of teachers is largely grade-specific rather than cohort-specific. In the initial 38 schools (and hence all of the 2013 data) we have two teachers in every school except one.

However, when restricting the students per classroom to be a minimum of 8 students we lose two teachers, leaving us with a total of 73 teachers in 2013.

In 2014, we have 122 new grade-one teachers from the 90 new schools and 22 new grade-one teachers from the original 38 schools that entered the sample. Of the teachers in the 2013 data, 44% are not present in the 2014 sample. When restricting the class size to a minimum of 8 students we lose 7 teachers, leaving us with a total of 178 grade-one teachers in 2013.

In 2015, 55% of the grade-one teachers in 2014 were still teaching grade one, 10% were teaching grade two or grade three and the remaining 35% were teaching higher grades or not found. In addition, two teachers from the 2013 sample re-entered and 16 new teachers entered the sample. Thus, in 2015 we have 148 grade-one teachers, 171 grade-two teachers and 46 grade-three teachers (in 2015 grade-three is only in the original 38 schools).

In 2016, 61% of the grade-one teachers in 2015 were still teaching grade one, 3% were teaching grade two or grade three and the remaining 37% were teaching higher grades or not found. In addition 31 teachers from 2013 and 2014 re-entered and 26 new teachers entered the sample, leaving us with 151 grade-one teachers. For grade two, 40% of the grade-two teachers in 2015 still taught grade two, and for grade three 37% still taught grade three. In total, we have 714 teachers across all years and grades; of these 274 (or 38%) we observe teaching in at least two years.

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 and 2016 this initial sample of grade-one students was retained, and tracked into grades two[7], three and four. In 2014, a new cohort of grade-one students was added to the study. Among this new cohort, 100 grade-one students were randomly selected from each of the 128 schools.[8] As with the first cohort, this cohort was also tracked into grades two and three in 2015 and 2016, respectively. In 2015, a third

[7] P2 in 2014 is not included due to lack of teacher information.

[8] 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.

and smaller cohort, 30 grade-one students randomly selected from each school, was added and tracked into grade two in 2016. In 2016, the fourth cohort was added, by randomly sampling 60 grade-one students in each school.

3.4 Data

Summary statistics for students and teachers are presented in Table 2. The average age across years and grades is around 9 years and approximately 50 percent of the students are girls. The baseline scores are slightly higher in the Randomized Teacher Sample compared to the Full and Longitudinal Sample. Moreover, the baseline scores in the Longitudinal Sample are slightly lower compared to the Full and Randomized Teacher Samples. Beside the baseline scores there are no differences in student or teacher characteristics between the four samples of teachers.

Table 2: Descriptive statistics

	Full Sample			**Longitudinal Sample**		
Students	*Control*	*Reduced-cost*	*Full-cost*	*Control*	*Reduced-cost*	*Full-cost*
Age	9.13	9.11	9.05	8.80	8.78	8.67
Female (%)	49.23	50.06	49.62	47.99	50.78	49.74
Baseline EGRA score	-0.23	-0.11	0.03	-0.30	-0.23	-0.16
Endline EGRA score	-0.06	0.16	0.41	-0.15	0.03	0.23
# Children	11,284	13,091	13,274	6,916	8,768	8,777
Teachers						
Age	42.92	44.31	41.18	42.66	44.64	40.95
Women (%)	55.25	44.00	39.91	58.81	41.18	41.65
Salary (shillings)	389,185	399,581	383,748	387,591	398,141	379,817
Experience	15.09	17.33	15.90	15.09	17.52	15.52
Years of education	14.01	13.82	13.96	13.95	13.79	13.96
Total score on Ravens Progressive Matrices	1.88	1.86	2.04	1.92	1.88	2.07
# Teachers	301	322	319	124	143	142

Table 2 (Cont.)

	Randomized Teachers Sample			Longitudinal Randomized Teachers Sample		
Students	*Control*	*Reduced-cost*	*Full-cost*	*Control*	*Reduced-cost*	*Full-cost*
Age	9.42	9.47	9.46	9.63	9.85	9.68
Female (%)	49.07	50.09	49.64	48.11	51.06	50.12
Baseline EGRA score	-0.18	0.00	0.20	-0.20	-0.01	0.16
Endline EGRA score	0.03	0.33	0.65	0.07	0.48	0.91
# Children	7,810	8,762	8,679	2,594	2,934	2,811
Teachers						
Age	42.21	44.11	41.00	41.99	42.36	39.90
Women (%)	55.24	42.33	40.17	52.67	54.36	54.05
Salary (shillings)	390,428	397,173	374,356	378,349	382,836	353,529
Experience	15.71	18.08	15.77	16.64	18.44	14.91
Years of education	13.98	13.88	13.96	14.07	14.43	14.14
Total score on Ravens Progressive Matrices	1.95	1.88	2.03	2.46	1.80	2.22
# Teachers	249	266	267	50	54	54

Notes: The Full Sample includes all teachers available in 128 study schools. The Longitudinal Sample includes all teachers who are teaching in at least two different years (from 2013-2016). The Randomized Teachers Sample includes all teachers teaching in 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 2016).

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 a principal components score index in the following way. First, we normalize each of the test modules against the control group, then we take the

(control-group normalized) first principal component as in Black and Smith (2006). This procedure is done separately for each year and grade.[9,10]

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.[11] 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 the beginning of 2013, 2014, 2015, and 2017. From these surveys we have information on both individual and household characteristics. We also conducted a three-question Raven's Standard Progressive Matrices (SPM) test to measure fluid intelligence, as well as asking a range of questions in social science, science, math and language.[12] 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, earns 390,000 shillings per month ($105), and has a total score of 2 out of 3 on the SPM test, or 66% correct. This score would put the average teacher at around the 50th percentile of the US adult distribution on the full 60-item SPM (Bilker et al. 2012). Roughly 43 percent are women.

In 2013 we also conducted in-person observations of each classroom in the study. These classroom observations were done by experienced enumerators and measured teacher and student 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.

[9] See Appendix B for the distributions of the endline PCA scores by grade level

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

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

[12] The SPM and general knowledge questions were only asked in 2013 and 2014.

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.

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:

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

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-supplied inputs ($\boldsymbol{S}_s(a)$) such as school management etc., cumulative

classroom inputs such as the teacher ($\boldsymbol{C}_{cs}(a)$) 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).

4.2 Empirical Strategy

Classroom Effects

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

$$Y_{icgst} = \beta_0 + \beta_1 Y_{icgst-1} + \boldsymbol{X}'_{icgst}\beta_2 + \lambda_{cgst} + \zeta_g + \beta_3 Y_{icgst-1}\zeta_g + \tau_t + \varepsilon_{icst} \quad (2)$$

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}).[13] $\boldsymbol{X}_{icgst}$ is a vector of individual characteristics and includes gender and age. 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 to vary with grade-level.

To estimate λ_{cgst}, three issues arise: First, there may be school effects that co-vary with true classroom effects, such as school management, resources or other factors that influence school choice. Second, there may be individual student effects that co-vary with true classroom effects, such as sorting of students to teachers based on parental influence or other unobserved characteristics. Third, sampling error: The estimated classroom effects are the sum of the true

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

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 (2) 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 $\hat{\lambda}_{cgst}$ to be relative to the school mean to thereby only consider the within-school variation in the classroom effects (Slater, Davies, and Burgess 2012, Araujo et al. 2016). The rescaled classroom effects become:

$$\hat{\gamma}_{cgst} = \hat{\lambda}_{cgst} - \frac{\sum_{c=1}^{C_s} N_{cs}\hat{\lambda}_{cgst}}{\sum_{c=1}^{C_s} N_{cs}} \quad (3)$$

$\hat{\gamma}_{cgst}$ is the demeaned classroom effect where C_s is the total number of classrooms we study within school s and N_{cs} is the total number of students we have test scores for in classroom c and school s. 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 Rothstein (2010), Kinsler (2012), Chetty, Friedman, and Rockoff (2014) and Goldhaber and Chaplin (2015) for a discussion of the severity of this bias). We address this potential source of bias by restricting our sample to the years (2013 and 2016) 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 attrition was correlated with teacher ability. We find no evidence of student switching or student attrition being systematically related to teacher characteristics (Appendix D).

Because we have years in our study when students were explicitly 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 compare the difference in 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 sampling variance. This is particularly problematic when we have a small number of student test scores in each class. To address this issue we take two approaches. First, we restrict the samples to only include classrooms with a minimum of eight students. Second, we analytically adjust the variance of the estimated classroom effects following the approach suggested by Araujo et al. (2016).[14] For the within-school classroom effects we estimate the variance of the measurement error as $\frac{1}{C}\sum_{c=1}^{C}\left\{\frac{\left[\left(\sum_{c=1}^{C_S} N_{cs}\right)-N_{cs}\right]}{N_{cs}\left(\sum_{c=1}^{C_S} N_{cs}\right)}\hat{\sigma}^2\right\}$, where $\hat{\sigma}^2$ is the variance of the residuals, ε_{icst} from equation (2). *C* is the overall number of classrooms in the sample. Then we subtract $\frac{1}{C}\sum_{c=1}^{C}\left\{\frac{\left[\left(\sum_{c=1}^{C_S} N_{cs}\right)-N_{cs}\right]}{N_{cs}\left(\sum_{c=1}^{C_S} N_{cs}\right)}\hat{\sigma}^2\right\}$ from the estimated variance of the demeaned classroom effects:

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

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

$$\hat{V}_{corrected}\left(\hat{\gamma}_{cgst}\right) = V\left(\hat{\gamma}_{cgst}\right) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{1}{N_{cs}}\hat{\sigma}^2\right\} \quad (6)$$

Teacher effects

The estimated classroom effects from equation (2) 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 and

[14] 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.

2016. We estimate teacher effects using the demeaned classroom effects with the following equation:

$$\hat{\gamma}_{cgst} = \hat{\alpha}_0 + \hat{\delta}_{cgs} + \omega_{cgst} \quad (7)$$

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. We correct for sampling variation 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:

$$\hat{\delta}_{cs} = \beta_0 + \boldsymbol{C}'_{cgs}\beta_1 + \psi_{cgs} \quad (8)$$

where $\hat{\delta}_{cgs}$ are our estimated teacher effects from equation (7), $\boldsymbol{C}_{cgs}$ is a vector of teacher characteristics and includes; gender, years of experience, salary, years of schooling and number of correct answers on the Raven's Standard Progressive Matrices (SPM).

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 time use, 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:

$$B_{blrcs} = \beta_0 + \beta_1\hat{\gamma}_{cgs} + \boldsymbol{C}'_{cgs}\beta_2 + \rho_s + \varsigma_r + \varphi_{rcs} + \omega_{blrcgs} + \mu_{lrcs} + \epsilon_{blrcs} \quad (9)$$

where *s* indexes schools, *c* indexes classrooms, *r* indexes the round of the visit, *l* indexes the lesson being observed, 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. $\boldsymbol{C}_{cgs}$ controls for teacher characteristics and includes: gender, experience, salary, years of schooling and number of correct answers on the SPM. 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. β_1 is our coefficient of interest and measures how classroom actions vary with teacher effectiveness.

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 (6) among all schools available (Pooled Sample) and columns 3 and 4 presents the estimates from the same equations using only schools in the control study arm (Control Sample). Columns 1 and 3 presents classroom value-added which is calculated using all teachers available (Full Sample) from equation (2) whereas columns 2 and 4 presents teacher value-added which is calculated using teachers with at least two years of data (Longitudinal Sample) using equation (6). Each of the estimates of classroom and teacher value-added measures are summarized 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.

The results in columns 1 and 2 can be interpreted as representing a setting in which school and teacher interventions are common, whereas the results in columns 3 and 4 can be interpreted as representing a setting with no school or teacher interventions. Overall, the value-added estimates are smaller when only using control schools.

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 0.24-0.35 SDs for the Pooled Sample (columns 1 and 2) and by 0.12-0.27 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).

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

	All Schools		**Control Schools**	
	(1)	(2)	(3)	(4)
Panel A: Including school effects	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
SD of effects	0.35	0.27	0.27	0.15
	[0.33-0.38]	[0.24-0.29]	[0.24-0.29]	[0.13-0.18]
Corrected SD of effects	0.32	0.24	0.23	0.12
	[0.29-0.35]	[0.21-0.27]	[0.21-0.26]	[0.09-0.15]
Panel B: School effects purged				
SD of effects	0.27	0.20	0.20	0.11
	[0.24-0.29]	[0.17-0.22]	[0.18-0.23]	[0.08-0.13]
Corrected SD of effects	0.23	0.17	0.17	0.07
	[0.20-0.26]	[0.14-0.20]	[0.14-0.20]	[0.04-0.10]
Children	37,649	24,461	11,284	6,916
Teachers	942	409	301	124
Schools	128	127	42	41
Pupils per classroom/teacher	27	31	26	30
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 2016. 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.

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 grade-level, 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 0.17 SDs and by 0.07 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 to the years where students were randomly assigned to teachers, in 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 0.21 SDs in the Pooled Sample (Column 1) and by 0.17 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 0.15 SDs in the Pooled Sample (Column 2) and by 0.03 SDs in the Control Sample (Column 4). Overall, we see that restricting the sample to random assignment years reduces the value-added estimates.

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

	All Schools		**Control Schools**	
	(1)	(2)	(3)	(4)
Panel A: Including school effects	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
SD of effects	0.37	0.31	0.27	0.17
	[0.33-0.39]	[0.27-0.35]	[0.21-0.33]	[0.13-0.21]
Corrected SD of effects	0.32	0.25	0.24	0.1
	[0.29-0.35]	[0.18-0.31]	[0.17-0.29]	[0.04-0.15]
Panel B: School effects purged				
SD of effects	0.26	0.21	0.21	0.12
	[0.21-0.30]	[0.17-0.25]	[0.15-0.27]	[0.07-0.15]
Corrected SD of effects	0.21	0.15	0.17	0.03
	[0.16-0.26]	[0.09-0.19]	[0.10-0.24]	[0.00-0.08]
Children	25,251	8,339	7,810	2,594
Teachers	782	158	249	50
Schools	128	103	42	33
Pupils per classroom/teacher	27	28	26	29
Sample	Full	Longitudinal	Full	Longitudinal

Notes: The Randomized Teachers 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.

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=255). 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 larger than the those under Business-as-usual, consistent with that higher quality teachers being matched with lower performing students.

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

	Classroom Effects		**Teacher Effects**	
	(1)	(2)	(3)	(4)
Panel A: Including school effects	Business-as-usual	Random assignment	Business-as-usual	Random assignment
SD of effects	0.21	0.33	0.21	0.21
Corrected SD of effects	[0.18-0.24]	[0.27-0.37]	[0.19-0.23]	[0.18-0.24]
	0.19	0.3	0.19	0.17
	[0.15-0.21]	[0.24-0.34]	[0.16-0.21]	[0.13-0.20]
Panel B: School effects purged				
SD of effects	0.16	0.27	0.16	0.16
	[0.13-0.18]	[0.23-0.32]	[0.14-0.18]	[0.14-0.18]
Corrected SD of effects	0.14	0.25	0.14	0.12
	[0.11-0.15]	[0.20-0.30]	[0.12-0.16]	[0.09-0.14]
Children	7,910	8,839	7,910	8,839
Teachers	255	255	255	255
Schools	122	122	122	122
Pupils per classroom/teacher	26	31	27	31
Sample				

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.

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). The fact that the classroom effect estimates are sensitive to the use of random assignment, while the teacher effect estimates are not, suggests that a substantial part of the systematic sorting of students into classrooms is not consistently occurring each year. This potentially makes the teacher effects a reasonable measure of teacher effectiveness, even in the absence of random assignment.

5.4 Robustness

In this section we address two issues: a) The imputation of grade-one baseline scores in 2015 and 2016 and b) compliance with random assignment in 2013, 2016 and 2017.

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 in 2013 and 2014 (in principle zeros). 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 more non-classical measurement error to our outcome variable and thus potentially bias 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 Full and Longitudinal Samples – essentially re-running the estimates for Table 3. Second, we replace all grade-one baseline scores with zero – including students in 2013 and 2014 for whom we have baseline test scores – and re-estimate the results again using the Full and Longitudinal Samples.

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

	Omitting Grade-One in 2016 and 2015		Replacing all Baseline Grade-One Scores with Zero	
	(1)	(2)	(3)	(4)
Panel A: Including school effects	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
SD of effects	0.43	0.38	0.36	0.27
	[0.37-0.48]	[0.35-0.39]	[0.33-0.37]	[0.25-0.28]
Corrected SD of effects	0.39	0.33	0.32	0.24
	[0.33-0.45]	[0.29-0.37]	[0.29-0.34]	[0.22-0.26]
Panel B: School effects purged				
SD of effects	0.25	0.15	0.27	0.20
	[0.22-0.27]	[0.14-0.17]	[0.25-0.28]	[0.15-0.19]
Corrected SD of effects	0.19	0.10	0.24	0.18
	[0.16-0.23]	[0.04-0.15]	[0.21-0.25]	[0.15-0.19]
Children	24,916	14,164	37,649	24,461
Teachers	795	336	942	409
Schools	127	123	128	127
Pupils per classroom/teacher	23.71	24.42	26.64	30.99
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 2016. 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.

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 0.10 SDs compared to 0.17 SDs in Table 3. 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 3. Thus, the decrease in columns 1 and 2 in Table 6 is more likely due to the change in sample size than the imputation of grade-one baseline scores. We therefore conclude that our results are unlikely to be sensitive to the imputation of grade-one baseline test scores.

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 10% of cases, which is not far from 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 yields similar results as in Table 4 and show 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.

Table 7: Robustness: Compliance with Random Assignment

	All Schools		**Control Schools**	
	(1)	(2)	(3)	(4)
Panel A: Including school effects	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Corrected SD of effects	0.36	0.3	0.27	0.17
	[0.32-0.38]	[0.26-0.34]	[0.21-0.33]	[0.12-0.22]
	0.31	0.24	0.24	0.11
	[0.27-0.34]	[0.16-0.30]	[0.17-0.30]	[0.02-0.18]
Panel B: School effects purged				
SD of effects	0.25	0.19	0.21	0.13
	[0.20-0.29]	[0.15-0.23]	[0.14-0.27]	[0.07-0.17]
Corrected SD of effects	0.2	0.13	0.17	0.05
	[0.14-0.25]	[0.07-0.18]	[0.10-0.24]	[0.00-0.12]
Children	23575	7153	7456	2302
Teachers	735	137	239	44
Schools	128	94	42	29
Pupils per classroom/teacher	27	28	27	29
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 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.

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 no obvious relationship between any of the teacher characteristics and our estimated teacher or classroom effects (Table 8).

Table 8: Correlation with Teacher Characteristics

	(1) *Teacher Effects*	(2) *Classroom Effects*
Years of Schooling	0.009	0.009
	(0.009)	(0.006)
Salary (log)	-0.146	-0.137
	(0.096)	(0.087)
Gender (1=Male)	0.029	0.016
	(0.029)	(0.023)
Experience (years)	0.002	0.006
	(0.008)	(0.006)
Experience^2 (years)	-0.000	-0.000
	(0.000)	(0.000)
Ravens Progressive Matrices	-0.004	-0.010
	(0.014)	(0.011)
General knowledge	0.002	0.001
	(0.002)	(0.002)
Observations	130	159
R-squared	0.038	0.058

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.

This finding – that effective teachers are difficult to identify *ex ante* through observed characteristics – is common in the literature and suggests that other measures are needed to identify the most-effective teachers (Azam and Kingdon 2015, Slater, Davies, and Burgess 2012, Araujo et al. 2016).

Next, we examine how classroom observation data correlate with teacher value-added, equation (12) in Tables 9 through 12. Table 9 shows the relationship between teacher effectiveness and time use and classroom management. Columns 1, 2 and 3 present the relationship between teacher effectiveness and three measures of teacher attendance and shows no significant relationship; note that all of the teachers were observed for their 30-minute lessons and we might not expect to observe teachers who spend less time on task more generally. Next, we investigate the relationship between teacher effectiveness and classroom management measured by the three combined factor indices of classroom management: “Keeps students focused”, “Solid lesson plan”

and “Active throughout classroom”. The results are presented in columns 4, 5 and 6 and show that the most-effective teachers have more structured and planned lessons.

Table 9: Teacher Behaviors: Time-use and Classroom Management

	(1)	(2)	(3)	(4)	(5)	(6)
	Time Use (minuttes)			**Clasroom Management**		
	Teaching	In Class not Teaching	Outside Class	Keeps Students Focused	Solid Lesson Plan	Active Throughout Classroom
Classroom Effects	0.001	-0.001	0.001	-0.272	0.077**	0.019
	(0.002)	(0.001)	(0.002)	(0.176)	(0.027)	(0.022)
Observation Windows	422	422	422	420	420	420
Adjusted R-Squared	.094	.06	.048	.099	.279	.174

Notes: Sample is observation windows, based on 145 individual lesson observations for 26 teachers in 16 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 10 focuses on 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 teacher effectiveness 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 and more time on sentences. Moreover, we also find that more-effective teachers are associated with a higher level of student participation. Panel B presents the results from estimating the relationship between teacher effectiveness and teaching methods and materials used. Here we find that more-effective teachers are more likely to have the individual students reading at the chalkboard. Moreover, we find no significant relationship between teacher effectiveness and materials used. However, the sign of the coefficients suggests that more-effective teachers are using the primers more and the chalkboard less.

Table 10: Classroom Observations: Reading Activities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A	**Element of Focus**				**Percent Pupils**		
	Sounds	Letters	Words	Sentences	**Participating**		
Classroom Effects	-0.042	-0.340***	-0.104**	0.198***	4.050*		
	(0.044)	(0.050)	(0.040)	(0.062)	(2.136)		
Observations	280	280	280	280	280		
Adjusted R-Squared	0.139	0.115	0.076	0.099	0.239		
Panel B	**Teaching Method**				**Materials Used**		
	Whole Class	Smaller Groups	Individual at Seat	Individual at Board	Board	Primer	Reader
Classroom Effects	0.080	0.022	0.115	0.190**	-0.101	0.056	0.000
	(0.083)	(0.034)	(0.067)	(0.070)	(0.074)	(0.068)	(0.057)
Observations	280	280	280	280	280	280	280
Adjusted R-Squared	0.045	0.14	0.046	0.054	0.179	0.207	0.255

Notes: Sample is observation windows, based on 145 individual lesson observations for 26 teachers in 16 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 11 considers the relationship between teacher effectiveness 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 similar results as in Table 10 namely that more-effective teachers are associated with students spending more time on sentences as well as more active students. In panel B we find that more effective teachers are associated with students spending more time on "air writing"[15] and copying text from the board, but less time on practicing handwriting. In addition, we find that more effective teachers have students using slates much more.

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

Table 11: Classroom Observations: Writing Activities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A			**Element of Focus**				**Percent Pupils**	
	Pictures	Letters	Words	Sentences	Name		**Participating**	
Classroom Effects	-0.087	-0.142	-0.003	0.249***	0.240***		9.832*	
	(0.136)	(0.115)	(0.109)	(0.051)	(0.075)		(5.477)	
Observations	169	169	169	169	169		169	
Adjusted R-Squared	0.049	0.122	0.267	0.34	0.353		0.161	

Panel B			**Teaching Method**			**Materials Used**	
	Air Writing	Handwriting Practice	Copy Text from Board	Writing own Text	Board	Slate	Paper
Classroom Effects	0.283***	-0.176**	0.254***	-0.061	-0.008	0.491***	-0.062
	(0.033)	(0.067)	(0.061)	(0.070)	(0.036)	(0.040)	(0.068)
Observations	169	169	169	169	169	169	169
Adjusted R-Squared	0.08	0.436	0.306	0.204	0.149	0.42	0.235

Notes: Sample is observation windows, based on 145 individual lesson observations for 26 teachers in 16 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 12 shows the association between teacher effectiveness and speaking/listening behaviors of the students. Having a more-effective teacher is associated with more student to teacher as well as student to student interactions.

Table 12: Classroom Observations: Pupils Speaking and Listening

	(1)	(2)	(3)	(4)	(5)
	To Partner	To Small Group	To Whole Class	To Teacher	Percent Pupils Participating
Classroom Effects	0.033	0.038*	-0.006	0.047**	1.986
	(0.062)	(0.019)	(0.035)	(0.017)	(1.407)
Observation Windows	411	411	411	411	411
Adjusted R-Squared	0.294	0.117	0.253	0.101	0.222

Notes: Sample is observation windows, based on 145 individual lesson observations for 26 teachers in 16 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.

In sum, we find that teacher effectiveness is positively correlated with more structured and planned lessons. Moreover, we find that when having a more-effective teacher students are participating more, interact more with the teacher and spend more time on sentences. However, these results should be interpreted as suggestive as teacher effectiveness could be correlated with unobserved teacher attributes which could themselves affect teacher behaviors. Nonetheless, these results provide a first step towards understanding who the good teachers are and what they do in the African context.

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 test 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 our classroom and teacher value-added estimates are affected by the introduction of the NULP. Table 13 presents the classroom value-added estimates using the Randomized Teachers Sample.

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

	Classroom Effects		
	(1)	(2)	(3)
Panel A: Including school effects	Control	Reduced-Cost	Full-Cost
SD of effects	0.27	0.38	0.43
	[0.21-0.33]	[0.25-0.50]	[0.36-0.49]
Corrected SD of effects	0.24	0.34	0.39
	[0.17-0.29]	[0.20-0.46]	[0.31-0.45]
Panel B: School effects purged			
SD of r effects	0.21	0.21	0.34
	[0.15-0.27]	[0.14-0.28]	[0.23-0.45]
Corrected SD of effects	0.17	0.15	0.29
	[0.10-0.24]	[0.06-0.22]	[0.17-0.42]
Children	7,810	8,762	8,679
Teachers	249	266	267
Schools	42	44	42
Pupils per classroom/teacher	26	27	27

Notes: The sample includes all teachers available. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.

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 8 reveal that the program greatly increases the variance of the classroom effects. Table 14 presents the teacher value-added estimates using the Random Longitudinal Sample.

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

	Teacher Effects		
	(1)	(2)	(3)
Panel A: Including school effects	Control	Reduced-Cost	Full-Cost
SD of effects	0.17	0.27	0.31
	[0.13-0.21]	[0.19-0.35]	[0.21-0.41]
Corrected SD of effects	0.09	0.19	0.23
	[0.04-0.15]	[0.08-0.30]	[0.08-0.36]
Panel B: School effects purged			
SD of effects	0.12	0.16	0.28
	[0.07-0.15]	[0.08-0.24]	[0.18-0.37]
Corrected SD of effects	0.03	0.05	0.20
	[0.00-0.08]	[0.00-0.15]	[0.08-0.32]
Children	2,594	2,934	2,811
Teachers	50	54	54
Schools	33	36	34
Pupils per classroom/teacher	29	29	28

Notes: The sample includes all teachers available. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.

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 are benefitting 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. This interpretation, however, requires that the rank of teachers is not affected by the NULP. 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 this assumption we follow Djerrari and Smith (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: Rank Preservation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Teacher Characteristics						
	Age	Ravens	Gender	Salary	Experience	Schooling	General knowledge
First quartile of TVA	2.184	0.735	-0.185	-0.070**	2.146	-0.140	0.235
	(2.810)	(0.564)	(0.155)	(0.034)	(2.253)	(0.380)	(4.313)
Second quartile of TVA	-3.679	0.428	0.040	0.027	-1.743	0.020	0.880
	(3.398)	(0.578)	(0.128)	(0.039)	(2.679)	(0.286)	(4.673)
Third quartile of TVA	-0.135	-1.017**	0.201*	0.091	0.242	0.381	0.717
	(2.883)	(0.402)	(0.117)	(0.057)	(2.443)	(0.331)	(5.270)
Fourth quartile of TVA	0.758	-0.351	-0.112	-0.030	-2.487	-0.018	-2.017
	(2.522)	(0.534)	(0.116)	(0.050)	(2.270)	(0.421)	(2.624)
Observations	264	141	325	320	296	321	141

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. The statistics in this table are the differences between full-cost and control group means. TVA = Teacher Value Added.

Each column represent a fixed teacher background variable (including age, ravens test, 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 estimate if there are significant differences in quartile means between the Full-cost program and the control group (corresponding to 4x7=28 test). Under the assumption of independence of the different tests, we would expect about two or three rejections. We obtain three rejections, thus just on the borderline of what we would expect at the 10% level. This provides suggestive evidence for consistency with the rank preservation assumption.

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 importance of observable characteristics for 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 teacher effectiveness by interacting teacher characteristics with indicators for teaching in a reduced-cost or full-cost program school.

The results in Table 16 show some evidence that more educated teachers and teachers with more general knowledge are benefitting more from the NULP. This is in line with the previous results that it is the best teachers that are benefitting the most.

Table 16: Effects of the NULP on the Relationship between Teacher Effectiveness and Teacher Characteristics

	Teacher Effects					
Experience (years)	-0.001					-0.001
	(0.005)					(0.012)
Reduced-cost Program*Experience	-0.001					-0.003
	(0.003)					(0.018)
Full-cost Program*Experience	-0.003					-0.006
	(0.003)					(0.017)
Years of schooling		0.006				-0.018
		(0.013)				(0.013)
Reduced-cost Program*Years of schooling		-0.004				0.022
		(0.015)				(0.016)
Full-cost Program*Years of schooling		0.011				0.066**
		(0.024)				(0.027)
Log salary (shillings)			-0.081			-0.002
			(0.064)			(0.079)
Reduced-cost Program*Log salary			0.109			-0.309*
			(0.117)			(0.170)
Full-cost Program*Log salary			0.397***			0.005
			(0.115)			(0.224)
Ravens score				0.035		0.025
				(0.023)		(0.016)
Reduced-cost Program*Ravens score				-0.027		-0.019
				(0.029)		(0.024)
Full-cost Program*Ravens score				-0.072**		-0.063**
				(0.034)		(0.030)
General knowledge					-0.000	0.001
					(0.002)	(0.002)
Reduced-cost Program*General knowledge					0.004	0.002
					(0.003)	(0.004)
Full-cost Program*General knowledge					0.008*	0.008*
					(0.004)	(0.004)
Observations	290	290	290	130	130	130
R-squared	0.257	0.257	0.278	0.177	0.158	0.269

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.

7. Conclusion

We use data from a randomized evaluation of a program delivering teacher training and support in northern Uganda to assess 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 0.09 to 0.19 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 0.09 standard deviations is strikingly 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 of "teachers matter" to information that would be useful for policy makers and administrators to recruit, train and support teachers it is important to know who the most effective teachers are and what they do in the classroom. To address this issue we correlated our estimated teacher effects with teacher characteristics and classroom behaviors. We found no evidence of currently observed teacher characteristics being associated with teacher effectiveness. However, we do find that more effective teachers are more likely to have a solid lesson plan and have more active students. This suggests that it is difficult to screen good teachers *ex ante*, but that designing personal policies based on *ex post* evaluation of teachers could be a way forward. 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.

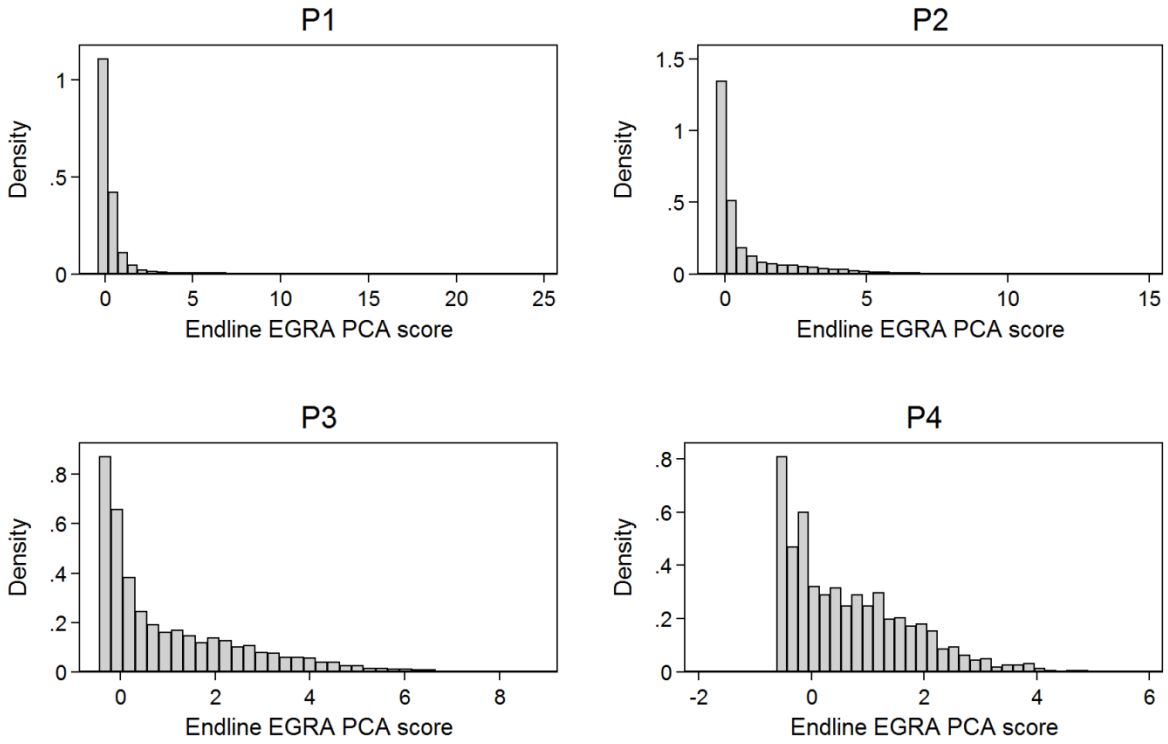
World Bank. *World Development Indicators 2010*, 2010 2010.

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

Appendices

Appendix B: Distributions of Endline PCA Scores by Grade Level

Figure B1: Distributions of Endline PCA Scores by Grade Level



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

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

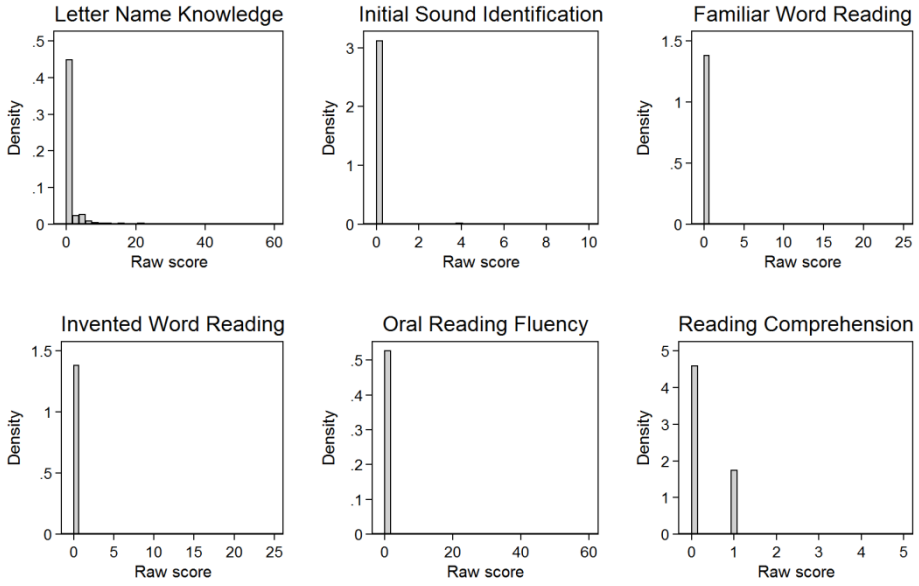
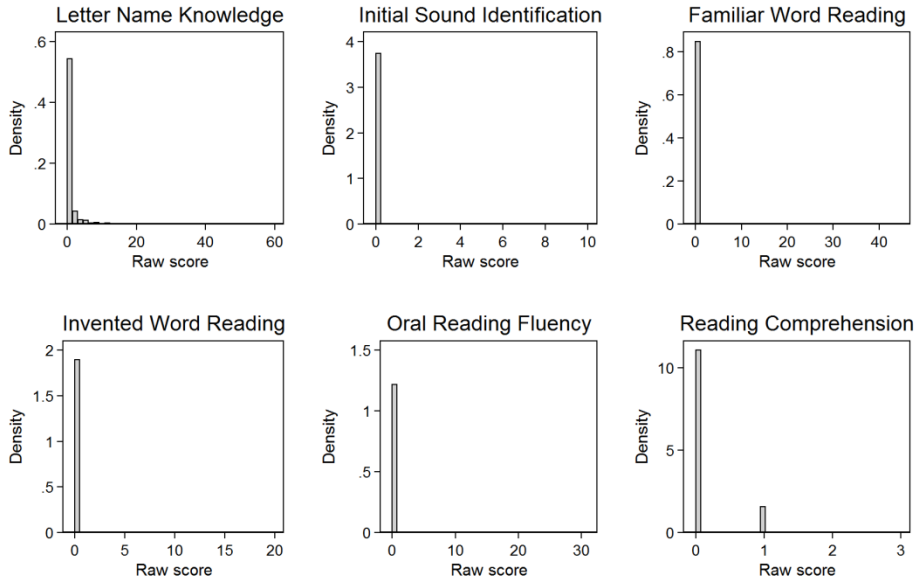


Figure C2: Distribution of the raw scores in the subtest for grade one in 2014



Appendix Table D: 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
Ravens score	.016
	(.019)
Observations	12517

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.

Appendix E Verifying Random Assignment

Figure E1: Distributions of P-values testing differences in baseline scores between classrooms within each school in 2013, 2016 and 2017

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

Figure E2: Distributions of P-values testing differences in baseline scores between classrooms within each school in 2014 and 2015

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

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about
our work on our website
www.theigc.org

For media or communications
enquiries, please contact
mail@theigc.org

Subscribe to our newsletter
and topic updates
www.theigc.org/newsletter

Follow us on Twitter
@the_igc

Contact us
International Growth Centre,
London School of Economic
and Political Science,
Houghton Street,
London WC2A 2AE

IGC
International
Growth Centre

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Designed by soapbox.co.uk