Teacher Effectiveness in Africa: Longitudinal and Causal Estimates

Julie Buhl-Wiggers, Copenhagen Business School

Jason T. Kerwin, University of Minnesota

Jeffrey A. Smith, University of Wisconsin and NBER

Rebecca Thornton, University of Illinois [*]

July 28, 2021

[*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 a 0.18 SD improvement in local-language reading and 0.20 in English reading at the end of one year. Teacher effectiveness is generally uncorrelated with observable teacher characteristics and we find no evidence of selective sorting of students to teachers. We then measure the causal effect of providing high-impact teacher training and support and find the program increases the variation in teacher effectiveness—by 61 percent in local language reading and 15 percent in English, most likely by improving teaching among the most-effective teachers.

---

[*] 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, the International Population Conference, NYU Steinhardt, Georgetown, UT Austin, Northeastern University, and the E-conomics of Education Seminar 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 and the Ichuli Institute, Katherine Pollman, Deborah Amuka and other Mango Tree Educational Enterprises staff. We are grateful for funding from DFID/ESRC Raising Learning Outcomes Grant ES/M004996/2, Wellspring, and the International Growth Centre.

1. Introduction

Teachers in sub-Saharan Africa face incredibly difficult working conditions. In northern Uganda, for example, they must deal with physical infrastructure that is in disrepair, classrooms with as many as 200 students, a lack of teaching materials, and insufficient training (Spreen and Knapczyk 2017). Over a tenth of teachers in Uganda do not have the formal qualifications needed to teach (Ugandan Ministry of Education and Sports, 2010); even those who are trained receive training that is of poor quality and inapplicable to the classroom (Hardman et al. 2011). As a result of these constraints, the quality of instruction is low across Africa: teachers are absent from the classroom nearly half the time, and teach for fewer than three hours a day; fewer than 10 percent of teachers meet minimum competency standards for language instruction or general pedagogy (Bold et al. 2017). The poor quality of instruction is reflected in learning outcomes: just 11 percent of fourth graders across the seven countries in their sample can read a simple paragraph.

How do these constraints affect the importance of teacher quality for student outcomes? An extensive literature from the US has shown that teacher value-added has large effects on children's success in school and in adulthood (Koedel et al. 2015). Recent research has shown that teacher quality is also very important in Latin America (Araujo et al. 2016) and south Asia (Bau and Das 2020). But there are no existing estimates of teacher value-added for sub-Saharan Africa. In principle, the difficult classroom conditions facing teachers in Africa could make teachers either more important or less important. On the one hand, the lack of resources may raise the premium to creativity and adaptability. On the other, in the limit as classrooms become sufficiently large, no student will learn anything no matter how talented their teacher.

This paper answers this question by constructing the first estimates of teacher value-added for an African country, using panel data from a randomized controlled trial in northern Uganda.

Our data comes from an evaluation of the Northern Uganda Literacy Program (NULP), a literacy program that was centered around providing intensive teacher training and support. The evaluation tracks students and teachers over the course of five years, from 2013 to 2017; students were randomized to classrooms in three of those years. Our main estimates focus on the control group from the evaluation. To construct estimates of teacher value-added, we regress end-of-year test scores on classroom fixed effects and a set of controls, most importantly prior test scores (Chetty, Friedman, and Rockoff 2014a). Our analysis focuses on characterizing the standard deviation of these estimates of teacher effectiveness. We correct for across-school sorting of students and teachers by rescaling the classroom effects to be relative to the school mean. To account for sampling error due to the finite sizes of classrooms, we apply the Araujo et al. (2016) analytic correction. We estimate teacher effects as the stable component of the classroom effects over time.

We show that despite extremely adverse working conditions, some teachers in Africa are highly effective. A one-standard deviation increase in teacher effectiveness improves local language reading test scores by at least 0.18 standard deviations and English reading by at least 0.20 standard deviations; teacher value-added is strongly correlated across subjects, with a correlation coefficient of 0.76. The standard deviation of teacher effects is about one third smaller than that of classroom effects in both subjects.

Teacher quality matters even more in Africa than it does in the United States. At 0.18 SDs our estimate of the standard deviation of teacher effects is nearly twice as large as the Chetty et al. (2014a) estimate for the effect of American primary-school teachers on native-language reading scores. Our estimates are also larger than those found for other developing countries: 0.09 for

Ecuador (Araujo et al. 2016) and 0.15 for Pakistan (from Bau and Das 2020).[2] Consistent with research from the US, we find little correlation between teacher value-added and attributes such as educational qualifications or experience.

Our results are essentially unchanged if we restrict our sample solely to the three years where students were randomly assigned to classrooms. To interpret estimates of teacher value added as a causal parameter—one that answers the question: "how does being assigned to a teacher of a particular effectiveness affect student achievement?"—we need to ensure that the estimates are not driven by students systematically sorting into classrooms (Rothstein 2010, Goldhaber and Chaplin 2012). Our estimates of classroom and teacher value-added are nearly the same when we compare the entire sample to the random assignment years. Moreover, when we examine the years with *status quo* classroom assignments, the distributions of student baseline test scores across classrooms within a given school and grade level are inconsistent with systematic sorting (Horvath 2015). Self-reports from headteacher surveys also suggest limited sorting of students to classrooms by ability.[3] Taken together, our results in Uganda are consistent with evidence from the US which finds little bias when comparing random assignment to status-quo student assignment (Kane and Staiger 2008).

These results are robust to a number of alternate choices about how we construct our sample and analyze the data. Restricting our analysis to just the longitudinal sample does not appreciably change our results. Increasing the minimum class size to 10 students also leaves our results unchanged; further increasing the minimum to 15 students does affect the results somewhat,

[2] Azam and Kingdon (2015) provide estimates from India that are substantially larger than ours, at 0.37 SDs, but differ in two key ways. First, their results are for gains over two years; this would correspond to an annual gain of roughly 0.18 SDs, which is similar to our result. Second, they focus on teachers in secondary, rather than primary, schools.
[3] A headteacher, sometimes called a headmaster, the equivalent of a principal in the US school system.

but only for the local-language reading teacher effects, and not for the other three estimates. Variations on the imputation rules we use for missing prior test scores barely change our findings. Finally, purging year-by-school effects (rather than overall school effects) also makes little difference for our main results.

We also provide the first causal evidence that teacher training can affect the variance of teacher effectiveness. The NULP program we analyze resulted in massive increases in student learning: after three years of the intervention, students in full-cost program schools score 1.35 SDs higher on local language reading tests and 0.73 SDs higher on English reading (Buhl-Wiggers et al. 2018). Students in reduced-cost program schools score 0.78 SDs higher on local language reading and 0.40 SDs higher in English reading.[4] By re-estimating the teacher value-added results for the two treatment arms, we show that both versions of the intervention increase the spread of the distribution of teacher effectiveness. In local-language reading, the SD of teacher value-added increases by 39 percent in the reduced-cost program and 61 percent in full-cost program schools; the latter effect is statistically significant at the 5 percent level. In English reading our estimates suggest a 25 percent and 15 percent increase in the SD of teacher value-added due to the reduced- and full-cost programs, respectively, although we cannot reject the null of equality across study arms.

This increase in the variance of teacher value-added is likely due to the program making the strongest teachers even better. We test for rank preservation by testing for the equality of teacher characteristics across the full-cost treatment and control groups within each quartile of value-added (Bitler et al. 2005, Djebbari and Smith 2008). We reject the null hypothesis of no difference in just

[4] Because the NULP focused on local-language reading, the effects on English suggest cross-subject spillovers. Other studies examining cross-subject spillovers include Aaronson, Barrow, and Sander (2007), Araujo et al. (2016), Buddin and Zamarro (2009), Koedel (2009), and Jackson (2012).

seven out of the 64 tests we conduct across the two test score variables, suggesting that the program had rank-preserving impacts on teacher effectiveness. Because the NULP increased the variance of teacher value-added, rank preservation suggests that the NULP achieved its impacts by improving teaching primarily among the most-effective teachers—helping the best teachers more than the worst teachers.

This paper is the first to unite two distinct literatures in economics on the connection between teaching quality and student learning, which employ two different approaches. The first uses student test scores to estimate teacher value-added. This literature has focused primarily on developed countries (see eg. Rivkin, Hanushek, and Kain 2005, Chetty et al. 2011, Chetty, Friedman, and Rockoff 2014a), although recent work has extended this research agenda to the developing world as well (Araujo et al. 2016; Azam & Kingdon 2015; Bau & Das 2020).[5] We contribute to this literature by providing the first estimates of teacher value-added from an African country.

The second 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 two literatures have accumulated evidence largely in separate spheres. Our study is the first to integrate these two approaches, and show how teacher quality is affected by an intervention aimed at improving student learning through teacher training.

[5] A related literature examines the value-added of schools rather than teachers. Three papers we know of estimate school value-added in developing countries: Crawfurd and Elks (2018) for Uganda, Blackmon (2017) for Tanzania, and Muñoz-Chereau and Thomas (2016) for Chile.

2. Setting and Intervention Details

2.1 Primary Education in Uganda

Primary education in Uganda consists of seven years of schooling (P1 to P7, corresponding to grades one through seven) starting at age six. Since 1997, primary school has officially been free of charge; however, as resources are scarce many schools still depend on contributions from parents. Still, the reform of 1997 was successful in getting children into school and the country's net enrollment rate is now above 90 percent (Deininger 2003, World Bank 2013).[6] Despite this relatively high level of access, late enrollment, repetition and early drop out remain major challenges throughout the country. Only about 60 percent of students transition from primary to secondary school (World Bank 2010).

Uganda faces major learning challenges in its primary schools. Bold et al. (2017) find that the vast majority (94 percent) of children in government primary schools could not read a simple paragraph, 54 percent could not order numbers correctly, 47 percent could not add double-digit numbers, and 76 percent could not subtract double-digit numbers. Even at the end of primary school, students have often learned very little: 15 percent of all P7 students leave primary school without mastering division, and 20 percent leave primary school without being able to read a short story (Uwezo 2016).[7]

2.2 Teachers in Africa and Uganda

Teachers in sub-Saharan Africa face severe constraints to their ability to teach effectively: they are undertrained, lack quality materials and methods for teaching, face crowded classrooms,

[6] Net enrollment is defined as the fraction of all primary school-age students enrolled in primary school.
[7] These statistics likely overstate student performance because schools discourage weaker students from attending in grade 7 in order to prepare the strongest students for the higher-stakes primary leaving exam (Gilligan et al. 2018).

and work in schools with nonexistent systems for tracking pupil performance and insufficient school supervision. Data from the Service Delivery Indicators (SDI) collected in Africa show that teachers spend limited time on task, and also lack the skills and knowledge to teach effectively when they are in the classroom: "essentially no public primary schools in … [Kenya, Mozambique, Nigeria, Senegal, Tanzania, Togo, and Uganda] offer adequate quality education" (Bold et al. 2017).[8] Uganda is no exception to these patterns: Bold et al. find that Ugandan teachers are absent from the classroom over 50 percent of the time, and spend just three of the scheduled seven hours a day on instruction. Just 16 percent of teachers in Uganda have the minimum knowledge needed to teach language classes, and only 4 percent meet minimum standards for general pedagogical training.

In Uganda, there are 11 different languages of instruction and in 2007, the government mandated local language instruction in lower primary (grades 1 to 3), shifting away from English. There are many obstacles to implementing this "mother-tongue first" policy, however, including underdeveloped orthographies, poor instructional methodologies for reading, and a lack of relevant and adequate reading materials in most of the languages of instruction. Moreover, the curricula for teacher training and primary education are not harmonized, the education system does not have

[8] The SDI data from schools in Africa show that effective teaching time amounts to only three hours a day with almost 60% of teachers being absent from the classroom. Just 16 percent of teachers have minimum knowledge in language and only four percent have minimum pedagogical knowledge. In regard to classroom practices, most teachers give positive feedback, but less than half 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. The data also suggest that teachers in Uganda are similar to other teachers in East Africa (Kenya and Tanzania) on knowledge tests, ability to prepare a lesson, and access to school supplies and infrastructure (Authors Calculations). On the other hand, they are more likely to be absent from school and less able to evaluate student progress/abilities than teachers in Kenya and Tanzania. In comparison to teachers in West and Southern Africa (Madagascar, Mozambique, Niger, Nigeria, and Togo), Ugandan teachers are more likely to be absent, but spend more of their time at the school actually teaching. Teachers in East Africa, including Uganda, score substantially higher on knowledge tests, are better able to prepare lessons, and have more access to supplies and better infrastructure than teachers in West and Southern Africa (Service Delivery Indicators, The World Bank; authors' calculations).

the capacity for effective monitoring of teacher performance (Ministry of Education and Sports, 2004).

Primary school teachers must obtain a certificate to teach in Uganda, requiring four years of secondary school followed by two years of pre-service teacher training. However, pre-service teacher education in Uganda is of poor quality and has limited applicability to the classroom (Hardman et al. 2011). An audit in 2010 found that 12.7 percent of primary school teachers did not have the correct qualifications to teach (Ugandan Ministry of Education and Sports, 2010). Teachers in Uganda receive in-service training referred to as Continuous Professional Development (CPD) which is intended to update competences required in the classroom. The CPD program is managed through primary teachers' colleges by Coordinating Center Tutors (CCTs). CCTs are typically recruited from experienced teachers and headteachers. They are responsible for providing workshops on Saturdays and during school holidays, and for school-based support such as conducting classroom observations and providing feedback to teachers and head teachers. However, CCTs receive limited training and support, making it difficult for them to effectively mentor teachers (Hardman et al. 2011).

2.3 The Northern Uganda Literacy Project (NULP)

The Northern Uganda Literacy Project (NULP) is an early-grade mother-tongue literacy program developed in response to the educational challenges facing northern Uganda. Almost half of the poorest 20 percent of Ugandans live in northern Uganda (Ministry of Finance 2003); the area has experienced decades of civil war leading to millions of internally displaced pleople, severe infrastructure shortages, large flows of refugees from South Sudan, and historical marginalization dating back to the early 20th century. This has resulted in an overstretched and poor-performing education system even relative to the rest of Uganda, with classrooms as large as 200 students,

limited educational materials, and limited support and training for teachers (Spreen and Knapczyk 2017).

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 provides residential teacher training throughout the school year and classroom support visits to give feedback to teachers. The program's approach 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 students, and slates, chalk, and wall clocks for first-grade classrooms.[9]

The full-cost NULP consisted of the original literacy program as designed by and delivered by Mango Tree and its staff. In addition, a reduced-cost NULP was implemented in some schools, following a "cascade" or "training-of-trainers" delivery model led by Ministry of Education coordinating center tutors (CCTs) rather than Mango Tree staff; teachers in these schools also received fewer support visits.[10]

The NULP was introduced to different grades during our study (Appendix Table 1, Panel A). In 2013 and 2014, first-grade classrooms and teachers received the NULP, in 2015 second-grade classrooms and teachers received the program, and in 2016, third-grade teachers received

[9] 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.

[10] Two of the material inputs provided by the NULP—the slates and wall clocks—were provided only to a subset of the schools in the reduced-cost version of the program.

the program.[11] 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 no longer received additional training or support visits.

3. Research Design

The NULP evaluation was conducted over five years, from 2013 to 2017. This section describes how schools, students, and teachers were sampled for the NULP evaluation. We then describe randomization, the data and samples used in our analyses, and balance and attrition.

3.1 NULP Evaluation: Sampling of Schools, Students, and Teachers

Schools

Schools were sampled for the NULP evaluation in two phases. In 2013, 38 eligible schools were selected to be part of the study. To be eligible, schools had to meet a set of criteria established by Mango Tree, the most important being that each school needed exactly two first-grade classrooms and teachers.[12] In 2014, 90 additional schools were added to the evaluation. The eligibility criteria for these new schools were less stringent with no minimum number of classrooms required.[13]

[11] In 2017, Mango Tree piloted a teacher mentor program with fourth-grade teachers in the reduced-cost and full-cost schools to provide support; no materials or pedagogical trainings or support were delivered. This intervention was much less intensive than the earlier years.

[12] The other eligibility criteria for 2013 were: desks and lockable cabinets for each P1 class, a student-to-teacher ratio in P1 to P3 of no more than 135 in 2012, being located less than 20 km from the headquarters of the coordinating centre, being accessible by road year round, having a head teacher regarded as "engaged" by the CCT, and not having previously received support from Mango Tree.

[13] The other eligibility criteria for 2014 were: having desks and blackboards in P1 to P3 classrooms and having a student-to-teacher ratio of no more than 150 students during the 2013 school year in P1 to P3.

Students

The NULP evaluation follows four cohorts of first-grade children who entered the study schools in 2013, 2014, 2015, and 2016, comprising a total sample of 27,943 students. Appendix Table 2 describes the sample of students in the study. In 2013, 50 first-grade students were randomly sampled from each of the 38 schools based on enrollment lists collected at the beginning of the school year (Cohort 1 baseline sample). An additional 30 second-grade students per school were added to this cohort at the 2014 endline (Cohort 1 endline sample). In 2014, 100 first-grade students were randomly selected from each of the 128 schools—sampled either at baseline or endline (Cohort 2).[14] In 2015, 30 first-grade students were randomly selected from each school (Cohort 3). In 2016, 60 first-grade students were randomly selected in each school (Cohort 4). All of the random samples of students were stratified by gender and classroom.

Teachers

Across the five years of the study, there were a total of 1,382 teachers who taught our sampled students (Table 1, Panel A). In Ugandan government primary schools, there is typically one teacher assigned to a given classroom, with multiple classrooms per grade. In our sample, on average, there are approximately two teachers per grade; this varies across year and school.

[14] The sampling procedure for cohort 2 differed slightly between the original 38 schools and the 90 schools added in 2014. In the 38 schools that participated in 2013, an initial sample of 40 grade one pupils was drawn at the 2014 baseline, and then 60 students were added at the 2014 endline following the same sampling procedure as at baseline. In the 90 new schools, the 80 students were selected at baseline with an additional 20 added at endline. The difference was due to the organizational difficulty of testing large numbers of students at baseline or endline at each school.

3.2 Randomization

Random assignment of NULP to schools

Schools in the study were each assigned to one of three study arms: 1) full-cost NULP, 2) reduced-cost NULP, and 3) control. Schools in the control group did not receive the NULP. Schools were grouped into stratification cells of three schools each.[15] Each stratification cell had its three schools randomly assigned to the three different study arms via a public lottery. In 2013 there were 12 full-cost treatment schools, 14 reduced-cost treatment schools, and 14 control schools. In 2014, 30 additional schools were added to each of the treatment arms for a total of 42 full-cost treatment, 44 reduced-cost treatment, and 44 control schools.

Random assignment of students to classrooms and teachers

Under the status quo, the assignment of students to classrooms in Uganda is specific to each school and depends on the approach used by the school's headteacher. Headteacher surveys conducted in 2017 asked about pupil assignment and find 18 percent of headteachers report sorting on student ability, 22 percent report sorting on student behavior, and 44 percent report trying to balance student gender; 14 percent and 15 percent report sorting based on parental influence or to keep friends together, respectively.[16] In three of the five years of the study (2013, 2016 and 2017), we explicitly instructed headteachers to randomly assign students to classrooms (Appendix Table

[15] The cells were formed by matching schools based on their coordinating centres (roughly equivalent to school districts), class sizes, number of classrooms, distance to coordinating centre, and primary leaving exam pass rate.
[16] Smaller numbers of head teachers also reported sorting students randomly, based on willingness to learn, height, disability, by gender of the teacher, by student age, and alphabetically.

1, Panel B).[17] In 2014 and 2015, head teachers were not given any guidance on how to assign students to classrooms.[18]

3.3 Data

We use three types of data: student test scores to measure learning outcomes, student characteristic data, and teacher characteristic data.

Learning Outcomes: Student Test Scores

Our student learning outcomes consist of test scores in both local language reading and English reading. Test administration varied somewhat by subject, year, and cohort, summarized in Appendix Table 1, Panel C. In 2013 and 2014, learning assessments were administered at the beginning and end of the school year, while in 2015, 2016 and 2017, learning assessments were administered only at the end of the year. In 2017, learning assessments were only administered among students in grades 3-5, meaning that Cohort 4 students were only assessed one year, when they were in grade one in 2016.

The exams were versions of the Early Grade Reading Assessment (EGRA), an internationally recognized exam that assesses early literacy skills (Dubeck and Gove 2015, Gove

[17] To randomize students to teachers, 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 by interviewing head teachers.

[18] Within a school, head teachers have discretion to assign teachers to specific grades. Our analysis does not account for sorting of teachers to particular types of schools or grades.

and Wetterberg 2011, RTI 2009, Piper 2010).[19] We used two different validated versions of the test—English and Leblango. For each language, we construct indices by first standardizing the separate exam components against the control group for each student-year-grade observation, and second, calculating the mean of the standardized components. These indices for local language and English reading are then standardized against the control group separately for each year and grade.

Because both government regulations and the NULP curriculum stipulate that first-grade students should only be exposed to local language reading and writing, English EGRA exams were conducted beginning in grade two; first-grade students were administered an oral English test.

Student Characteristics

Our student-level analyses control for student age and student gender. In addition we control for ability using prior-year test scores on local language reading, English reading as well as math. The math score, is based on several questions that measured numerical pattern recognition, one- and two-digit addition and subtraction, and matching numbers to objects. Math tests were self-administered while led by facilitators in a group and are also standardized to the control group for each year and grade.[20]

Teacher Characteristics

[19] Both versions of the EGRA that we use cover six components of literacy skills: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. The English-language EGRA also has a letter sounds module.

[20] Given that the intervention focused on literacy, we do not report estimates of teacher value-added for math; these estimates. Math was assessed at the same times as local language reading, with the exception of not being collected at baseline in 2013.

Teacher characteristics come from teacher surveys and employee rosters. Teacher surveys were conducted in 2013 (Grade 1 teachers), 2014 (Grade 1 teachers), 2015 (Grades 1-3), and 2017 (Grades 3-5). Rosters of current and prior employees were collected from each school in 2014-2017. From these surveys, we have information on each teacher's age, gender, years of experience teaching, and years of education. We use 2015 levels for time-varying variables such as age and years of experience.

3.4 Construction of Analytical Samples

Annual Student Learning Gains

Our analytical strategy involves measuring the gain in student learning attributable to a teacher in a given school year. Appendix Table 3 provides a detailed description of the tests used to estimate teacher value-added for each subject, grade, and year of the study.

For each student, we need an endline test score in any given year; we drop student-year observations in which a student is missing endline tests in both local language and English reading. This results in 58,777 student-year observations and 27,943 unique students (Table 1, Panel A).

Next, for every student-year observation with an endline test, we identify their prior performance. To do so, we either use a student's endline assessment from the previous year, or, for grade-one students, we assign them a baseline score of zero.[21]

Because first grade students were not tested in English reading, we estimate English reading value-added only for students in grades two and above. This also implies that we do not

[21] This is motivated by the fact that 1) we only have baseline exams for a small subset of students in grade 1 in 2013 and 2014 and 2) among the students who were assessed at the beginning of the first grade, the majority (83%), scored zero on their local language reading test. Our results are unaffected if instead we focus only on students with baseline exams, or only impute exams that are missing.

include Cohort 4 students in the English analysis because they were not assessed in 2017 when they were in grade two. For students in grade two, we use oral English scores from the previous year while for students in grades three, four, and five, we use previous year English reading score to estimate learning gains.

In some cases we have an endline test score for a student, but are missing a prior test score, if, for example, a student was absent on the day of the exam. In that case, we impute students' missing prior test score as the median baseline score for her class. For both local-language and English reading, around 10,000 student-year observations have missing prior test scores; rates of missingness do not vary significantly by study arm (Table 1, Panel B). To account for student-year observations with missing prior scores, we include a dummy variable in our analysis indicating whether the prior score was missing. We also perform additional robustness checks (described in Section 5.4 below) to address missing prior scores.

Main Analytical Samples: Two-Teacher and Longitudinal Samples

To estimate classroom and teacher value-added we match students to specific teachers using classroom registers and student reports. Across 58,777 total students-year observations for which we have at least one endline exam, we are able to match almost all to a teacher (99 percent); this rate does not vary systematically across year or treatment arm (99.4 percent in the full-cost treatment, 98.7 percent in the reduced-cost treatment, and 99.1 percent in the control group). The most common reasons for not being able to match students to teachers include missing or misreported teacher names.[22] To limit estimation error due to sampling variation, we drop student-

[22] Misreported teacher names can lead mechanically to a teacher appearing to have only a single student, because only one student misreported the name in that way. The majority of teachers with such small numbers of students are likely to be artifacts of the data and not actual teachers, or in some cases, are teachers of students who have repeated a grade.

year observations from very small classrooms (fewer than five students per teacher in a given year). This removes 2,193 observations, corresponding to 3.8 percent of the overall sample bringing us down to 56,032 student-year observations (Table 1, Panel B). Again, the rate of very small classrooms does not vary much across randomization years or across school treatment status (2.9 percent in the full-cost treatment, 3.8 percent in the reduced-cost treatment, and 4.7 percent in the control).

To estimate *classroom* value-added, we need at least two teachers in each school to purge out school effects. Because we follow the same schools over time, we could purge either overall school effects or year-specific school effects. The fact we have fewer classrooms per school in the earlier years of the intervention (a new cohort was added each year) means that we also have systematically fewer teachers per school in earlier years. This means that purging year specific school effects will drop relatively more teachers from earlier years as we have more schools with only one teacher, which would limit our ability to draw comparisons of teacher value-added over time. To avoid this, we purge overall school effects instead of year-specific school effects. Table 1, Panel C shows the number of schools and teachers meeting this criterion, forming our *Two-Teacher Sample*: 56,032 student-year observations (27,608 unique students) and 1,763 classrooms (1,096 unique teachers).

To separate *teacher* value-added from *classroom* value-added, we need to observe a teacher over multiple years. We observe 475 (43 percent) of teachers teaching at least two years (41 percent in the full-cost treatment, 44 percent in reduced-cost treatment, and 47 percent in the control). This is our *Longitudinal Sample* and includes 1,138 classrooms (475 unique teachers) and 38,078 student-year observations (24,217 unique students). See Table 1, Panel D.

Teacher Characteristics Sample

Table 1, Panels C and D present the number of teachers for whom we have teacher characteristic data. Of the 1,096 unique teachers in our Two-Teacher Sample, we have teacher characteristics for 871 teachers (80 percent): 280 in the full-cost program (78 percent), 306 in the reduced-cost program (81 percent), and 281 in the control group (81 percent). Of the 475 teachers in the Longitudinal Sample, we have characteristics for 433, or 91 percent; 130 in the full-cost program (89 percent), 154 in the reduced-cost program (93 percent) and 149 in the control (91 percent).

3.5 Balance and Attrition

Balance across NULP Treatment Arms

Appendix Table 4 presents descriptive statistics for students and teachers in each of our analytical samples, separated by study arm. Schools are generally balanced across study arms in terms of student characteristics—age and gender—and teacher characteristics. Half of students are female (recall that the sample is stratified by gender), and students are almost nine years old (Panel A). On average, teachers are almost 40 years old, 48 percent female, with 14 years of education and 14 years of experience (Panel B).

Balance Tests for Random Assignment of Students to Teachers

To assess the degree of compliance with the random assignment of students to classes in 2013, 2016 and 2017 we perform two checks. First, we test if teacher characteristics are orthogonal to student characteristics, which gives us an indication of whether certain students are matched to certain teachers. Appendix Table 5 presents regressions of student characteristics on teacher

characteristics. While there are some significant coefficients, the majority are small and insignificant.

A second check for balance across randomly assigned students to teachers, we test the difference in student characteristics between teachers within schools and grade levels for each year, which indicates the degree of sorting similar students into the same classes (Horvath 2015). Appendix Figure 1 presents a distribution of *p*-values from regressing baseline test scores on teacher dummies within each year, school and grade-level. Looking at differences between baseline classroom test scores, we find that between 1 to 5 percent of the schools had classrooms with statistically-significant baseline differences between across classroom streams at the 5 percent level in 2013, 2016 and 2017. Overall, we can reject baseline balance in just 3.7 percent of cases, which is exactly what we would expect by random chance.

Student and Teacher Attrition

Student attrition from the study could be due to dropping out, transferring to another school, or being absent for an exam. The extent to which certain types of students attrit—either overall or differentially by study arm—could affect our external and internal validity of our analysis. In general, attriters tend to be older and girls are less likely to attrit; otherwise we do not see any concerning differences in student attrition across study arms.[23]

[23] Online Appendix Table 1 presents the correlation between student characteristics and student attrition—defined as a missing student-year observation of test scores—and examines attrition by study arm. Two threats to the validity of the value-added approach would be if students systematically switched classrooms during the year, or if student dropout was correlated with teacher ability. Online Appendix Table 2 presents the correlation between teacher characteristics and student attrition and shows that students with a female teacher are more likely to attrit in the reduced- and full-cost NULP study arms but not in the control group.

Teacher attrition is an important issue, given that our Longitudinal Sample requires observing a teacher over at least two years. We find that female teachers are less likely to attrit, however, this does not vary between study arms.[24]

4. Empirical Strategy

The "Value-Added Model" takes prior student achievement into account to control for variation in initial conditions (Rivkin, Hanushek, and Kain 2005; Todd and Wolpin 2003) and is typically an estimate of the increase in learning attributable to a specific classroom or teacher.[25] This section describes our empirical approach to estimating classroom and teacher value added and the causal effects of the NULP.

4.1 Classroom Effects

We begin by estimating classroom effects using the following "lagged-score" value-added model, separately for local language reading and English reading:[26]

[24] See Online Appendix Table 3. One caveat is that we observe characteristics for a only subset of teachers (Table 1).

[25] A canonical model of learning can be written as $Y_{icgsa} = Y_a\left[X_{icgs}(a), S_s(a), C_{cgs}(a), \theta_{i0}, \varepsilon_{icgsa}\right]$, where Y_{icsa} is a measure of achievement for child i in classroom c, grade g, and school s at age a. Acquisition of knowledge is a combination of cumulative family-supplied inputs, $X_i(a)$, cumulative school-level inputs, $S_s(a)$, such as school management, cumulative classroom inputs such as teacher ability, $C_{cgs}(a)$, and genetic endowments, θ_{i0}. ε_{icgsa} allows for measurement error in achievement. To estimate teacher value-added, student achievement can be modeled linearly with $Y_{icst} = \beta_0 + \beta_1 Y_{icst-1} + \beta_2 X_{icst} + \rho_s + \lambda_{cs} + \varepsilon_{icst}$, where Y_{icst} is a measure of achievement for child I in classroom c, grade g, and school s in year t. Prior achievement, 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 or school infrastructure. λ_{cs} is the effect of being in a specific classroom and an estimate of the increase in learning due to a specific classroom or teacher. $Var(\lambda_{cs})$, is interpreted as the variation in teacher effectiveness.

[26] 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, in which we do not control for lagged test scores and instead replace the left-hand-side of Equation (3) with $\Delta Y_{icgst} = Y_{icgst} - Y_{icgst-1}$.

$$Y_{icgst} = \beta_1 Y_{i,t-1} + \beta_2 Z_{i,t-1} + \beta_2 X_{icgst} + \lambda_{cgst} + \zeta_g + \beta_3 Y_{i,t-1} * \zeta_g + \beta_4 Z_{i,t-1} * \zeta_g + \beta_5 D_{icgst} + \beta_6 ST_{icgst} + \varepsilon_{icst} \quad (1)$$

where Y_{icgst} is the endline test score (Leblango or English) for child *i* in classroom *c*, in grade *g*, in school *s*, in year *t*. $Y_{icgst-1}$ is the student's prior test score. $Z_{icgst-1}$ is a vector of prior scores for the other reading exam and math. Because the predictive power of the prior test scores increases sharply with grade level—recall that the vast majority of children score zero in grade one—we let the effect of prior scores differ by grade level β_3 and β_4 are the grade specific effects of prior test scores. X_{icgst} is a vector of individual characteristics, specifically gender and age. λ_{cgst} is classroom fixed effects; year fixed effects are implicitly controlled for by the classroom fixed effect. We include (expected) grade-level (ζ_g) fixed effects as some students are repeaters and thus expected grade-levels could vary within each classroom. We dummy out any missing values of prior test scores, age, and gender by controlling for the vector of dummy variables D_{icgst}, a vector of dummies indicating that an observation was originally missing. Finally, we include an indicator for the sample type ST_{icgst}, which is equal to one if the child was sampled at endline and zero for students in the baseline sample. To estimate a full set of classroom effects, we omit the constant term from the regression.

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*. To estimate λ_{cgst}, three issues arise: First, there may be school effects or school-level shocks that co-vary with true classroom effects due to factors such as school management or school quality. Second, there may be individual student effects that co-vary with true classroom effects due to sorting of students to

teachers based unobserved characteristics. Third, 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 per teacher. As the sample gets smaller (fewer students tested per class) the sampling error increases. This sampling error could cause 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.

Purging school effects from classroom effect estimates

When estimating Equation (1) 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 classroom effects. Since students were randomized to classrooms only within schools, and not across them, some of the variation in our estimated classroom effects is in fact due to across-school sorting of students. To overcome this issue we rescale the classroom effects ($\hat{\lambda}_{cgst}$) to be relative to the school mean of the estimated classroom effects 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)):

$$\hat{\gamma}_{cgst} = \hat{\lambda}_{cgst} - \hat{\lambda}_{s} \qquad (2)$$

This approach nets out (in expectation) all school-level factors and thereby provides a lower bound on the degree of variation in the classroom effects, since some of the across-school variation in classroom effects represents real differences in teaching quality.

Sorting of students to teachers

Endogenous sorting of students to teachers can introduce bias to value-added estimates (Rothstein 2010; Kinsler 2012; Chetty, Friedman, and Rockoff 2014; Goldhaber and Chaplin 2015). Because we have some years of data where students were randomly assigned to teachers, for a subset of our overall sample of teachers we can test the null hypothesis that the variances of the classroom or teacher effects are equal under random assignment. Specifically, we compare the random-assignment years to all years for the same set of teachers, to get a sense of the severity of the bias due to sorting.

Sampling variance

The estimated variance of the classroom effects is the sum of the true variance and the sampling variance. The latter term arises because the classroom effects are estimated with a finite sample of students. The smaller the number of students, the more likely that the estimated effect on learning of a given classroom will be large due to random chance. Thus this issue is a particular concern when we have a small number of student test scores in each class. To address this issue we follow the approach suggested by Araujo et al. (2016).[27] For the within-school classroom effects, we estimate the variance of the measurement error and subtract that from the estimated variance of the de-meaned classroom effects:[28]

$$\hat{V}_{corrected}(\hat{\gamma}_{cgst}) = V(\hat{\gamma}_{cgst}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{\left[\left(\sum_{c=1}^{C_s} N_{cs}\right) - N_{cs}\right]}{N_{cs}\left(\sum_{c=1}^{C_s} N_{cs}\right)}\hat{\sigma}^2\right\} \quad (3)$$

[27] The procedure is analogous to an 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 de-meaned within each school, and that the within-school mean may also be estimated with error. See online appendix D of Araujo et al. (2016) for details.

[28] This reduces to $\hat{V}_{corrected}(\hat{\gamma}_{cgst}) = V(\hat{\gamma}_{cgst}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{1}{N_{cs}}\hat{\sigma}^2\right\}$ when using both between- and within-school variation to estimate classroom effects.

where $\hat{\sigma}^2$ is the variance of the residuals, ε_{icst}, from Equation 1. C is the overall number of classrooms in the sample, and N_{cs} is the number of students in classroom c in school s. Motivated by concerns about sampling variation, our main specification uses only data from classrooms with at least five students. As robustness check we also restrict the sample to only classes with a minimum of 10 or 15 students, excluding smaller classrooms.

4.2 Teacher effects

The estimated classroom effects from Equation (1) contain both a permanent teacher component as well as a transitory classroom component that captures disturbances during testing or peer dynamics during a particular year. Because classroom effects include both teacher effects and random classroom shocks, classroom effects will have a higher variance than teacher effects. When we have more than one year of data for the same teacher it is possible to separate teacher effects from classroom effects, under certain assumptions. The identifying assumption is that ω_{cgst} is not serially correlated across years.

We estimate teacher effects using the classroom effects with the following equation:

$$\hat{\gamma}_{cgst} = \hat{\delta}_{cgs} + \omega_{cgst} \quad (4)$$

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 teacher effects. With this approach, we assume that all time variation in the classroom effects is due to transitory shocks and not changes in actual teacher quality. If this assumption fails, ω_{cgst} could contain "real" teacher quality fluctuations, and our teacher effects estimates, $\hat{\delta}_{cgs}$, would be biased toward zero. We de-

mean our estimated teacher effects by the school average to purge any school effects.[29] We correct the variance of the teacher effects for sampling variation, in the same manner as described above when estimating classroom effects, see Equation 3.

4.3 Value-Added Correlations with Teacher Characteristics

To examine if teacher characteristics can explain variation in our estimated measure of teacher effectiveness we estimate the following equation:

$$\hat{\delta}_{cgs} = \beta_0 + C'_{cgs}\beta_1 + \psi_{cgs} \quad (5)$$

where $\hat{\delta}_{cgs}$ are our estimated teacher effects from Equation (4), C_{cgs} is a vector of teacher characteristics and includes gender, years of experience, and education level.[30]

5. Estimates of Classroom and Teacher Effectiveness under the Status Quo

5.1 Main Estimates

Table 2 presents our estimates of teacher and classroom effects using the two-teacher and longitudinal samples. We present the results among students in control schools only to understand how teacher value-added is distributed under the status quo, without the NULP intervention. To summarize the distributions of the various classroom and teacher value-added estimates, we present the standard deviation of each estimate, measured in terms of standard deviations of student performance on the end-of-year exams. We present our estimates with and without corrections for sampling variance and present cluster-bootstrapped confidence intervals in square brackets.

[29] $\hat{\zeta}_{cgs} = \hat{\delta}_{cgs} - \hat{\delta}_s$

[30] We convert all time-varying variables (i.e. age and experience) to their 2015 levels for comparability.

Panel A shows the results for local-language reading. Columns 1 and 2 use both between- and within-school variation to estimate classroom and teacher effects, and indicate substantial variation across classrooms and teachers. After correcting for sampling variation, a one-SD increase in classroom quality increases student performance in local-language reading by 0.30 SDs; for teacher effects, the estimate is 0.21 SDs (Panel A, Columns 1 and 2). Because the estimates in Columns 1 and 2 also include between-school variation, some proportion of the estimated variation is likely to be due to non-random sorting of teachers and students to schools. By implication, these estimates are upper bounds on the variance of the true γ_{cgst} (classroom effects) and δ_{cgs} (teacher effects).

To purge the variation of school-level effects, in Columns 3 and 4 we limit our analysis to within-school variation only, effectively comparing teachers between classes in the same year and school. Using this specification, we still find substantial variation between teachers, although smaller magnitudes. The estimated variance of teaching quality for local-language reading is slightly smaller, with our preferred estimate showing that a one SD increase in classroom/teacher quality is associated with an increase in student performance by 0.29/0.18 SDs.

To put the differences between the first two columns and the second two columns into context, it is useful to consider two extreme possibilities in terms of how much teachers sort into schools based on their effectiveness. If there is no sorting at all, then the estimates without school effects measure the true variance of teacher value-added in the entire population of teachers. If teachers were perfectly sorted to schools with the most-effective teachers working together in one school, and the least effective in one school as well, then the estimated variance of teacher value-added after removing school effects will approach zero. In intermediate cases, the estimates with school effects purged serve as a lower bound on the overall variance of teacher effectiveness.

Panel B shows the analogous results for English reading. Here, the estimates including school effects are somewhat larger at 0.48/0.40 SDs (Panel B, Columns 1 and 2). After purging the school effects, the estimates are between 35 and 46 percent smaller; our preferred estimated classroom/teacher value-added are 0.31/0.20 SDs (Panel B, Columns 3 and 4).

Local language teacher value added is highly correlated with English: the two teacher effect estimates for the two subjects (after purging school effects) have a correlation coefficient of 0.76. This estimate is attenuated relative to the true correlation due to the estimation error in constructing the two value-added estimates (Goldhaber, Cowan, and Walch 2013). It is similar to the correlations typically found in developed countries, which range from 0.20 to 0.80, with most results being toward the higher end of that scale (Koedel and Betts 2007, Loeb, Kalogrides, and Beteille 2012, Goldhaber, Cowen, and Walch 2013, Loeb, Soland, Fox, and Kun 2015, and Condie, Lefgren, and Sims 2014).

5.2 Random Assignment of Students to Classrooms

To investigate the degree of bias due to sorting of students to classes, we re-estimate the teacher and classroom effects using a different sample of teachers—those who teach in the random assignment years, 2013, 2016, and 2017. Columns 1 and 2 present the results from all years, for the subset of teachers for whom we have data in years of random assignment. Columns 3 and 4 restrict the sample to these teachers, during 2013, 2016, and 2017; this enables us to compare estimates using all years of data to those that use only the random assignment years for the same set of teachers.

Table 3 presents the results for local language reading (Panel A) and English (Panel B). In both cases, the difference in variance of classroom and teacher effects across years with random

student to classroom assignment is negligible, as compared to the estimates that use the full sample. The fact that our estimates do not vary greatly across assignment regimes is in line with the evidence from head teacher surveys and balance tests in student characteristics. We conduct tests for sorting in the non-randomized years that parallel the tests for sorting in the random-assignment years discussed above and presented in Appendix Figure 1 (Horvath 2015). Only 1 to 5 percent of the schools had classrooms with statistically significant baseline differences between streams.

5.3 Correlations between Teacher Value-Added and Teacher Characteristics

Using data from the teacher surveys, we describe how teacher characteristics correlate with value-added estimates in Table 4. Except for having a bachelor's degree—which is negatively associated with value-added—we find few patterns of predictors of value-added.[31] In general, the predictive power of teacher characteristics for teacher value-added is quite low: at the high end, our covariates can explain less than one percent of the variance in value-added.

5.4 Sensitivity Analysis

We present several different robustness tests for our main estimates of value-added from Table 2. We address issues related to: a) the sample composition of teachers, b) conditioning on a particular minimum classroom size, c) the construction of learning gains when baseline or prior-year test scores are missing, and d) purging school-year effects rather than overall school effects.

First, because the teachers in the two-teacher sample and the longitudinal sample differ somewhat from each other (56% of teachers in the two-teacher sample are not in the longitudinal

[31] Zakharov et al. (2016) find that teacher age and educational credentials correlate with student performance in South Africa.

sample), Appendix Table 6 presents the equivalent estimates of classroom and teacher effects conditioning on teachers being in the longitudinal sample. The results are similar to those in Table 2, which means that teacher attrition does not seem to invalidate our main results.

Next, because our data contain a (random) sample of students within each classroom, in some cases we have a rather small number of students per teacher.[32] Our preferred estimates in Table 2 condition on teachers having at least five students in their classroom. As the statistical consistency of the value-added estimates depends on the number of students per teacher, we assess the sensitivity of the inclusion of small class sizes on our results by re-estimating our results from Table 2, omitting class sizes either below 10 or below 15 students per teacher. Appendix Table 7 shows that excluding classrooms with fewer than 10 or 15 sampled students per teacher barely changes the estimated variance of classroom effects. It does increase the variance of the teacher effects, but only for local-language reading.

We next address the fact that we impute missing prior test scores to avoid losing student-year observations. Appendix Table 8, Columns 1 and 2 presents the estimates without imputing the prior scores—in other words, we omit any student-year observation without a prior test scores. Panel A presents the results for local language reading and Panel B presents the results for English reading. The variances of the classroom and teacher effects are only slightly affected relative to those in Table 2.

Because baseline tests were not administered in 2015 and 2016, first-grade students that were recruited into the study in those years have no prior exam scores available. Thus the estimates in Table 2 involve imputing grade-one baseline test scores to zero, which (for consistency) we do

[32] As noted above, many of the small "classrooms" are likely to be misreporting of teacher names by a small number of students, including misspellings of names as well as the use of nicknames.

for all first-grade student. Columns 3 and 4 of Appendix Table 8, Panel A present results for instead omitting all first-grade students from the estimates.

As a final sensitivity test we present the results when purging year-specific school effects as opposed to overall school effects in Appendix Table 9. Again we find that the variance of teacher effects barely changes for either local language reading (Panel A) or English (Panel B). However, the variance of classroom effects is smaller for both subjects, in comparison to the estimates in Table 2.

6. Effects of the NULP on the Distribution of Teacher Value-Added

6.1 Classroom and Teacher Value-Added

While previous research 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 distribution of teacher value-added when teachers are provided with comprehensive training and support. Recall that the NULP program was highly effective: an analysis of the effects of the program suggests massive effects on learning, with local-language reading scores increasing by 1.35 standard deviations in the full-cost program and 0.78 standard deviations in the reduced-cost version, after three years of exposure to teachers in the program (Buhl-Wiggers, et al. 2018).

In Table 5, we show how the introduction of the NULP affects the variance of our classroom (Columns 1-3) and teacher (Columns 4-6) effect estimates. Columns 1 and 4 show the results for teachers in schools that did not get the program, and so simply replicate the results in Columns 1 and 2 in Table 2. Columns 2 and 5 present the results for reduced-cost program schools and Columns 3 and 6 the results for the full-cost program schools.

In both local-language reading and English reading, the program increases the variance of classroom and teacher effects. The corrected standard deviation of classroom effects increases by 21 and 38 percent in Leblango reduced- and full-cost program schools, and by 3 and 6 percent in English reduced- and full-cost program schools. The estimated increases in the standard deviation of teacher effects due to the program are even larger: 39 and 61 percent for local language and 25 and 15 for English, reduced- and full-cost, respectively. Based on assessing whether the bootstrapped confidence intervals for the estimates overlap (and thus assuming the covariance term is zero), we can reject the null hypothesis that the local-language reading classroom and teacher effects are equally distributed in the control group and the full-cost program schools. We cannot reject the null of equal standard deviations for the reduced-cost program schools and the control group, nor for any of the comparisons for English reading. Given that the program's main emphasis was on local language reading, the increases in English suggest that language teaching ability translates across languages, so that teachers whose Leblango teaching ability was improved by the program also saw their English teaching ability improve. This conclusion is subject to two limitations. First, due to limited statistical power, we cannot rule out the equality of the distributions of teacher value-added for English. Second, the program does lead to average reading gains in English, because it does provide some training and inputs for English reading, particularly at the third-grade level (Buhl-Wiggers et al. 2018).

6.2 Testing for Rank Preservation

The finding that a highly effective teacher-training program increases the spread of teacher effectiveness in Leblango and English means that some teachers improve more than others. Since the program leads to gains in student performance on average in those subjects, the most intuitive

explanation is that the impact of the program was largest for the highest-quality teachers. It seems unlikely that the program would have made skilled teachers perform worse, which would be needed in order for it to sharply alter the rankings of teacher ability. A very strict version of this interpretation requires rank preservation. This means that, for example, a teacher at the median of the value-added distribution in the full-cost program should also have as her counterfactual the median teacher in the control-group distribution. To test an implication of the rank preservation assumption we follow (Bitler et al. 2006; Djebbari and Smith 2008) and test whether fixed covariates have the same means in a given quantile of the teacher value-added distribution. We focus on comparisons of the full-cost program and control-group schools; our results are similar when we compare the reduced-cost program schools to the control group.

Table 6 presents the results of tests for rank preservation. Each column represent a fixed teacher background variable (age, gender, experience and degree obtained). Each row corresponds 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 4x4=16 tests). Under the (incorrect) assumption of independence of the different tests, we would expect about two or three rejections. For Leblango, we obtain two rejections when using the classroom effect estimates and zero when using the teacher effect estimates. For English, there is one rejection for the classroom effects and four for the teacher effects. We thus reject the null at the 10 percent level in 7 out of the 64 total tests, or 10.9 percent of the time. Our evidence is therefore consistent with the theory that the treatment had rank-preserving effects on teacher value-added. There are three caveats to these results. First, we do not have characteristics on all our teachers, so we cannot test this using the full sample of teachers. Second, the power of this test is limited by the fact that teacher characteristics are only

weakly correlated with teacher effects. Thus our failure to reject the null may simply reflect low power. Third, even a high-powered version of this test is one-sided in nature: if the test rejects the null hypothesis, then we know that the rankings of the teachers were shifted by the treatment, but it is possible for the rankings to be affected without altering the quartile-specific distributions of the covariates—for example, if teachers are re-sorted only within quartiles and not across them.

6.3 Sensitivity Analysis

As described above, the NULP intervention was only implemented for certain grade levels in certain years; see Appendix Table 1. To address sensitivity to this feature we perform two sensitivity tests that are presented in Appendix Table 10. First, we omit data collected in 2017 as the NULP was only implemented from 2013 to 2016 (Columns 1-6). This leaves the classroom effect estimates nearly unchanged, but reduces the estimated variance for the teacher effects; this difference may be because we are effectively putting more weight on lower grades. This change does not change our conclusion that the NULP increased the variance of teacher value-added. Second, we restrict our sample to only include teachers teaching in classes directly affected by the NULP for the two treatment groups, and the corresponding teachers in the control group (Columns 7-12). These estimates of show similar patterns to Table 5.

7. Conclusion

We use data from a randomized evaluation of a multi-faceted literacy program that focuses heavily on 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. Our estimates are credible: utilizing the fact that students were randomly assigned to teachers, we test across assignment regimes for bias due to sorting of students to teachers; we show that sorting is not an issue for estimation in this setting. Second, we show that in this context there is no evidence that formal education or teaching experience are important determinants of teacher value-added. Third, we are able to shed light on how a high-impact teacher-training program, the NULP, affects the spread of the teacher quality distribution.

Our results show that under the *status-quo* (i.e., in the control group from our experiment) a one standard deviation increase in teacher effectiveness increases student learning by at least 0.18 SDs. However, unlike in developed countries, because the "best" performing teachers are unlikely to be at the frontier of the education production possibilities, this result does not necessarily indicate how much teachers can improve. In this paper we directly test how teacher training and support as provided by the NULP affects the variance of the teacher value-added estimates. We find that the NULP increases the spread of the teacher value-added distribution, making teachers more diverse in their effect on student learning. Our data is consistent with rank preservation, which implies that the program achieved its large average gain by improving the performance of the strongest teachers while leaving the weaker ones behind. This result suggests that an important avenue for future research is to look at how to better reach the less-effective teachers.

A caveat is that our findings are limited somewhat by our sample size: our estimated confidence intervals do not allow us to reject the null hypothesis that some of the patterns we document could have arisen through random chance. However, we can reject the null of equally-distributed teacher effects for local-language reading in the control group and in the full-cost

NULP. Sample size limitations also restrict our ability to dig into certain patterns that are interesting, such as year-by-year differences in classroom effects, and means that our power to examine correlations between teacher effects and teacher characteristics is fairly low.

Direct evidence on teaching quality in Africa is scant. Such evidence is needed: despite near universal enrollment in primary school, students in Africa are lagging behind the rest of the world in competencies in foundational skills. Our study provides evidence that effective teacher training and support can indeed increase teacher value-added even in an extremely low-resource context. We also find that it is possible to achieve large improvements in teacher value-added through teacher training. Unfortunately, the NULP helps not the worst-performing teachers but the best ones. This means that teacher training interventions may have limited utility in terms of improving the effectiveness of lower-performing teachers.

Moreover, observed teacher characteristics only explain a small fraction of the variance in teacher value-added, and thus *ex ante* screening of teachers based on traditional measures such as education levels and experience may be difficult. More research is needed on how to design policies based on *ex post* evaluation of teachers, and on whether there are alternative characteristics that predict teacher effectiveness *ex ante*. Solving the learning crisis in Africa will require novel ideas for helping improve the quality of teaching across the entire distribution of teacher performance.

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. " Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." Journal of Economic Perspectives—Volume 31, Number 4—Fall 2017—Pages 185–204 .

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.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, 40, 76–92. https://doi.org/10.1016/j.econedurev.2013.11.009

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.

Gilligan et al. 2018

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.

Ministry of Education and Sports' (2003- 2004) Curriculum Review

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.

Spreen, C., & Knapczyk, J. J. (2017). Measuring Quality Beyond Test Scores: The Impact of Regional Context on Curriculum Implementation (in Northern Uganda). FIRE: Forum for International Research in Education, 4(1). Retrieved from http://preserve.lehigh.edu/fire/vol4/iss1/1

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.

Wills, Gabrielle. 2017. "What do you mean by 'good'? The search for exceptional primary schools in South Africa's no-fee school system," Working Papers 16/2017, Stellenbosch University, Department of Economics.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, 100, 61–78. https://doi.org/10.1016/j.jpubeco.2013.01.006

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

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

World Bank. 2015. Report No: AUS3446. Education Service Delivery in Nigeria. Results of 2013 Service Delivery Indicator Survey. October 2015

Zakharov, Andrey, Gaelebale Tsheko, Martin Carnoy. 2016. "Do "better" teachers and classroom resources improve student achievement? A causal comparative approach in Kenya, South Africa, and Swaziland." *International Journal of Educational Development*, Volume 50, Pages 108-124, https://doi.org/10.1016/j.ijedudev.2016.07.001.

Tables and Figures

Table 1: Samples Across Study Arms

Panel A: NULP Evaluation Sample	All	Control	Reduced Cost	Full Cost
#Schools	128	42	44	42
#Teachers	1,382	470	485	427
#Classrooms	2,200	728	762	710
#Sampled students	29,790	9,573	10,456	9,761
#Students with at least one EL test	27,943	8,948	9,799	9,196
Panel B: Students with Consecutive Tests				
#Student-year obs with EL local language tests	58,777	18,638	20,421	19,718
#Student-year obs with prior & EL local language tests	49,053	15,425	17,042	16,586
#Student-year obs with EL English tests	37,059	11,712	12,815	12,532
#Student-year obs with prior & EL English tests	27,270	8,485	9,402	9,383
Panel C: Matching Students to Teachers				
#Student-year obs matched to a teacher	58,225	18,478	20,157	19,590
#Student-year obs in a classes larger than 5 students	56,032	17,612	19,391	19,029
Panel D: Two-Teacher Sample				
#Schools	128	42	44	42
#Teachers	1,096	365	384	347
#Teachers with data on characteristics	871	282	308	281
#Classrooms	1,763	568	614	581
#Students	27,608	8,820	9,670	9,118
#Student-year obs	56,032	17,612	19,391	19,029
Panel E: Longitudinal Sample				
#Schools	125	40	44	41
#Teachers	475	146	167	162
#Teachers with data on characteristics	435	132	154	149
#Classrooms	1,138	347	397	394
#Students	24,217	7,468	8,678	8,071
#Student-year obs	38,078	11,430	13,280	13,368

Notes: The 128 schools where sampled in two phases: 38 in 2013 and additional 90 in 2014. Prior test is defined as an EL test in the year before. The Two-Teacher Sample includes all students and 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 and their students. Both the Two-Teacher Sample and the Longitudinal Sample are based on students with local language test, the numbers for English are smaller.

Table 2: Classroom and Teacher Value-Added: Two-Teacher and Longitudinal Samples, Control Schools

	Including School Effects		School Effects Purged	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Panel A: Leblango Reading				
SD of effects	0.32	0.24	0.30	0.20
	[0.25,0.39]	[0.17,0.31]	[0.24,0.36]	[0.15,0.25]
Corrected SD of effects	0.30	0.21	0.29	0.18
	[0.22,0.37]	[0.14,0.29]	[0.22,0.35]	[0.12,0.24]
Observations (student-years)	17,612	11,430	17,612	11,430
Students	8,820	7,468	8,820	7,468
Teachers	365	146	365	146
Classrooms	568	347	568	347
Schools	42	40	42	40
Pupils per classroom/teacher	43	78	43	78
Panel B: English Reading				
SD of effects	0.49	0.41	0.32	0.22
	[0.32,0.66]	[0.25,0.56]	[0.27,0.37]	[0.18,0.26]
Corrected SD of effects	0.48	0.40	0.31	0.20
	[0.31,0.66]	[0.24,0.55]	[0.25,0.37]	[0.16,0.24]
Observations (student-years)	10,882	6,116	10,882	6,116
Students	5,675	4,360	5,675	4,360
Teachers	284	99	284	99
Classrooms	390	211	390	211
Schools	42	40	42	40
Pupils per classroom/teacher	37	55	37	55

Notes: Classroom effects are calculated from the Two-Teacher sample of teachers and teacher effects are calculated from the Longotudinal sample of teachers. The Two-Teacher 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 school mean. Control schools (N=42) did not receive the NULP intervention.

Table 3: Comparison with Random Assignment Value-Added Estimates using the Same Sample of Teachers

	All Years		Random Assignment Years	
Panel A: Leblango Reading	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Corrected SD of effects	0.30	0.20	0.30	0.19
	[0.24,0.36]	[0.12,0.28]	[0.24,0.35]	[0.11,0.27]
Observations (student-years)	14,705	3,144	10,393	2,131
Students	8,359	2,384	7,349	1,888
Teachers	285	39	285	39
Classrooms	478	106	346	78
Schools	42	13	42	13
Pupils per classroom/teacher	42	81	39	55
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal

	All Years		Random Assignment Years	
Panel B: English Reading	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Corrected SD of effects	0.30	0.10	0.27	0.10
	[0.25,0.35]	[0.06,0.14]	[0.24,0.30]	[0,0.26]
Observations (student-years)	9,093	1,722	7,766	1,464
Students	5,289	1,430	4,985	1,289
Teachers	216	29	216	28
Classrooms	317	68	266	57
Schools	42	13	42	13
Pupils per classroom/teacher	37	58	38	50
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal

Notes: Panel A includes all years (2013-2017) but conditioning on teachers teaching in random assignment years. Panel B includes only random assignment years (2013, 2016 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. All estimates are purged of school effects by subtracting off the school mean. Control schools (N=42) did not receive the NULP intervention.

Table 4: Teacher Value-Add Correlation with Teacher Characteristics, Control schools

	Leblango Reading		English Reading	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
≥ Bachelor (1=Yes)	-0.075**	-0.051	-0.089*	-0.015
	(0.035)	(0.035)	(0.049)	(0.064)
Female (1=Yes)	-0.036	-0.004	-0.057	0.005
	(0.033)	(0.041)	(0.039)	(0.047)
< 5 yrs of experience (1=Yes)	-0.006	0.064	0.073	0.250
	(0.140)	(0.235)	(0.136)	(0.272)
yrs of experience	-0.000	-0.001	0.006*	0.007
	(0.003)	(0.003)	(0.004)	(0.005)
< 5 yrs of experience (1=Yes)*	0.013	-0.013	-0.004	-0.032
yrs of experience	(0.045)	(0.073)	(0.045)	(0.092)
Observations	470	132	310	87
R-squared	0.017	0.014	0.034	0.048
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal

Notes: Standard errors are clustered by school, in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The dependent variables are teacher and classroom effects.

Table 5: Heterogeneity of Value-Added by NULP Study Arm, School Effects Purged

	Classroom Effects			Teacher Effects		
Panel A: Leblango EGRA	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD of effects	0.29	0.35	0.40	0.18	0.25	0.29
	[0.22,0.35]	[0.30,0.41]	[0.37,0.44]	[0.11,0.24]	[0.18,0.32]	[0.24,0.34]
Observations (student-years)	17,612	19,391	19,029	11,430	13,280	13,368
Students	8,820	9,670	9,118	7,468	8,678	8,071
Teachers	365	384	347	146	167	162
Classrooms	568	614	581	347	397	394
Schools	42	44	42	40	44	41
Pupils per classroom/teacher	31	32	33	78	80	83
Sample	Two-Teacher			Longitudinal		

	Classroom Effects			Teacher Effects		
Panel B: English EGRA	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD of effects	0.31	0.32	0.33	0.20	0.25	0.23
	[0.25,0.37]	[0.25,0.39]	[0.29,0.37]	[0.16,0.24]	[0.16,0.33]	[0.17,0.29]
Observations (student-years)	10,882	11,943	11,945	6,116	6,975	7,213
Students	5,675	6,130	5,973	4,360	4,911	4,995
Teachers	284	297	277	99	100	110
Classrooms	390	416	389	211	233	228
Schools	42	44	42	40	44	41
Pupils per classroom/teacher	28	29	31	55	58	61
Sample	Two-Teacher			Longitudinal		

Notes: The sample includes the Two-teacher Sample for classroom effects and the Longitudinal Sample for teacher effects. All estimates are purged of school effects by subtracting off the school mean. 95% confidence intervals for the SD of the classroom effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.

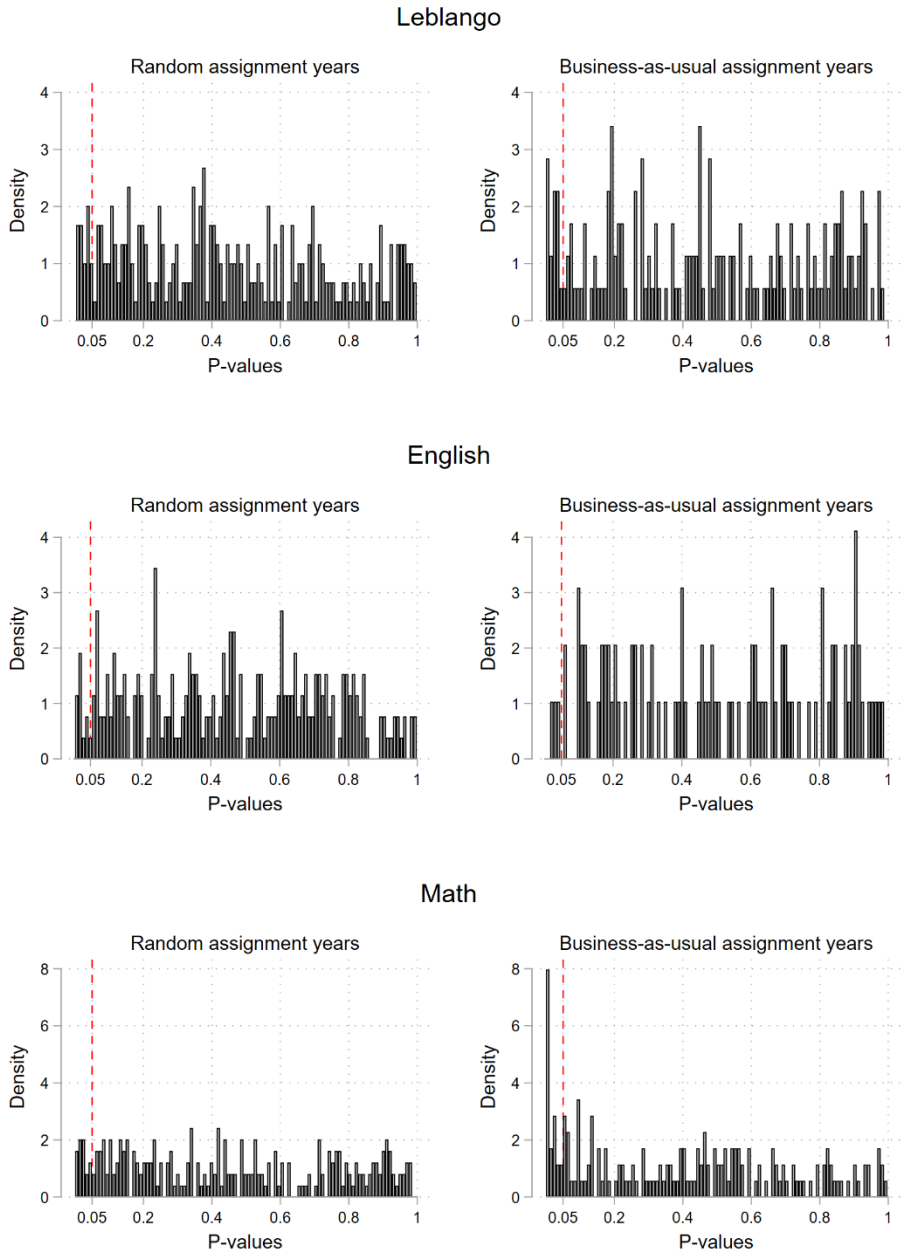
Table 6: Tests of Rank Preservation

	Dependent Variable: Difference between Full-Cost and Control in Teacher Characteristic							
	Classroom effects				Teacher Effects			
Panel A: Leblango Reading	Age	Gender	Experience	Schooling	Age	Gender	Experience	Schooling
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First quartile of TVA	-2.142	0.055	-1.838	-0.011	-2.006	0.128	-1.558	0.025
	[-2.582,2.306]	[-0.131,0.144]	[-2.400,2.298]	[-0.106,0.111]	[-3.048,3.351]	[-0.204,0.202]	[-3.031,3.143]	[-0.123,0.122]
Second quartile of TVA	2.542	-0.139*	2.624*	-0.023	2.647	-0.008	-0.140	-0.082
	[-2.553,2.761]	[-0.129,0.140]	[-2.643,2.551]	[-0.105,0.112]	[-2.996,2.966]	[-0.216,0.205]	[-3.188,3.221]	[-0.201,0.174]
Third quartile of TVA	-0.358	0.065	-0.740	-0.011	-0.753	-0.058	1.539	0.040
	[-2.474,2.572]	[-0.146,0.145]	[-2.276,2.132]	[-0.106,0.115]	[-3.759,4.043]	[-0.227,0.206]	[-3.424,3.392]	[-0.156,0.151]
Fourth quartile of TVA	-0.360	-0.005	-0.779	0.036	-0.041	-0.087	-2.123	0.061
	[-2.177,1.993]	[-0.134,0.127]	[-2.032,2.093]	[-0.108,0.113]	[-3.068,3.058]	[-0.174,0.170]	[-3.184,3.192]	[-0.120,0.118]
Observations	569	600	563	600	284	291	281	291
	Classroom effects				Teacher Effects			
Panel B: English Reading	Age	Gender	Experience	Schooling	Age	Gender	Experience	Schooling
	(1)	(2)	(4)	(5)	(6)	(7)	(9)	(10)
First quartile of TVA	1.391	-0.016	1.754	-0.061	-0.227	0.203	3.260	0.076
	[-2.892,2.826]	[-0.146,0.149]	[-2.697,2.678]	[-0.150,0.141]	[-3.382,3.362]	[-0.254,0.270]	[-3.768,3.883]	[-0.196,0.166]
Second quartile of TVA	1.265	-0.126	1.952	0.054	4.079*	-0.265*	4.055**	-0.153
	[-2.593,2.549]	[-0.148,0.152]	[-2.644,2.393]	[-0.116,0.121]	[-3.611,3.620]	[-0.273,0.260]	[-3.495,3.288]	[-0.181,0.175]
Third quartile of TVA	-1.541	0.089	-1.577	-0.053	-3.158	-0.115	-3.147	0.034
	[-2.680,2.515]	[-0.139,0.148]	[-2.347,2.225]	[-0.118,0.120]	[-4.142,4.148]	[-0.207,0.217]	[-3.761,3.969]	[-0.182,0.175]
Fourth quartile of TVA	-1.278	0.003	-3.221**	0.008	-2.164	0.122	-5.519**	0.037
	[-2.260,2.374]	[-0.152,0.157]	[-2.380,2.563]	[-0.129,0.125]	[-3.674,3.735]	[-0.235,0.221]	[-3.706,3.708]	[-0.188,0.195]
Observations	447	473	437	473	188	194	184	194

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 Longitudinal Sample).

Appendix

Appendix Figure 1
Distributions of P-values testing differences in baseline scores between classrooms within each school in random assignment year



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 Table 1: NULP Treatment, Student Assignment to Classroom and Assessment by Year

Panel A: NULP Treatment	2013 (1)	2014 (2)	2015 (3)	2016 (4)	2017 (5)
Grade treated	Grade 1	Grade 1	Grade 2	Grade 3	Grade 4
Panel B: Student Assignment to Classrooms					
Random assignment of students to classrooms?	Yes	No	No	Yes	Yes
Panel C: Learning Assessments Administered	2013 (1)	2014 (2)	2015 (3)	2016 (4)	2017 (5)
Grades assesed	Grade 1	Grades 1-2	Grades 1-3	Grades 1-4	Grades 3-5
Leblango reading tests (all grades)	Baseline and Endline	Baseline and Endline	Endline	Endline	Endline
English oral tests (grade-one only)	Baseline and Endline	Baseline and Endline	Endline	Endline	--
English reading tests (grades > 1)	--	Baseline and Endline	Endline	Endline	Endline

Appendix Table 2: Number of Students per School Sampled by School Sample and Year

Panel A: Original 38 schools sampled in 2013

	2013	2014	2015	2016
Cohort 1 baseline sample	50 grade-1 students			
Cohort 1 endline sample		30 grade-2 students		
Cohort 2 baseline sample		40 grade-1 students		
Cohort 2 endline sample		60 grade-1 students		
Cohort 3			30 grade-1 students	
Cohort 4				60 grade-1 students

Panel B: New 90 schools sampled in 2014

	2013	2014	2015	2016
Cohort 2 baseline sample		80 grade-1 students		
Cohort 2 endline sample		20 grade-1 students		
Cohort 3			30 grade-1 students	
Cohort 4				60 grade-1 students

Appendix Table 3: Tests Used to Estimate Value-Added

Panel A: Leblango Reading

	2013		2014		2015		2016		2017	
Cohort 1	Grade 1	0, EL 2013	Grade 2	EL 2013, EL 2014	Grade 3	EL 2014, EL 2015	Grade 4	EL 2015, EL 2016	Grade 5	EL 2016, EL 2017
Cohort 2			Grade 1	0, EL 2014	Grade 2	EL 2014, EL 2015	Grade 3	EL 2015, EL 2016	Grade 4	EL 2016, EL 2017
Cohort 3					Grade 1	0, EL 2015	Grade 2	EL 2015, EL 2016	Grade 3	EL 2016, EL 2017
Cohort 4							Grade 1	0, EL 2016	Cohort not Assessed	

Panel B: English Reading

	2013		2014		2015		2016		2017	
Cohort 1	Grade 1	Not assessed in English reading	Grade 2	EL oral English 2013, EL 2014	Grade 3	EL 2014, EL 2015	Grade 4	EL 2015, EL 2016	Grade 5	EL 2016, EL 2017
Cohort 2			Grade 1	Not assessed in English reading	Grade 2	EL oral English 2014, EL 2015	Grade 3	EL 2015, EL 2016	Grade 4	EL 2016, EL 2017
Cohort 3					Grade 1	Not assessed in English reading	Grade 2	EL oral English 2015, EL 2016	Grade 3	EL 2016, EL 2017
Cohort 4							Grade 1	Not assessed in English reading	Cohort not Assessed	

Appendix Table 4: Descriptive Statistics across Treatment Arms and Samples

	Two-Teacher Sample			Longitudinal Two-Teacher Sample		
Panel A: Students	Control	Reduced Cost	Full Cost	Control	Reduced Cost	Full Cost
	(1)	(2)	(3)	(4)	(5)	(6)
Female (%)	0.497	0.508	0.495	0.497	0.507	0.496
Age	8.897	8.932	8.949	8.953	8.990	8.989
Panel B: Teachers						
Age	0.457	0.440	0.391	0.464	0.457	0.405
Female (%)	39.784	40.330	39.716	39.794	40.562	39.509
Yrs of experience	14.070	14.202	14.371	14.257	14.490	14.146
<5 yrs of experience	0.101	0.098	0.104	0.093	0.092	0.105
Yrs of education	14.799	14.610	14.629	14.750	14.598	14.590
Teachers college or below	0.282	0.345	0.322	0.298	0.353	0.331
Diploma	0.515	0.494	0.511	0.510	0.491	0.517
Bachelor or above	0.202	0.161	0.167	0.192	0.156	0.152
#Teachers with characteristics data	280	306	281	130	154	149

Notes: The full sample includes all teachers available. The two teacher sample includes schools with at least two teachers in a given year (2013-2017). The longitudinal sample includes all teachers who are teaching in at least two different years (from 2013-2017).

Appendix Table 5: Correlation between Student and Teacher Characteristics

	Student Characteristic							
Dependent variable:	Female				Age			
Treatment arm	Control	Reduced-cost	Full-cost	All	Control	Reduced-cost	Full-cost	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.011	0.010	0.011	0.010**	-0.005	0.000	0.010	0.006
	(0.009)	(0.009)	(0.008)	(0.005)	(0.028)	(0.033)	(0.023)	(0.015)
Age	0.001*	0.000	0.000	0.001	0.007**	-0.001	0.006**	0.004**
	(0.001)	(0.001)	(0.001)	(0.000)	(0.003)	(0.002)	(0.003)	(0.002)
Bachelor or above	0.007	-0.003	-0.010	-0.002	-0.028	0.018	-0.032	-0.006
	(0.011)	(0.010)	(0.013)	(0.006)	(0.036)	(0.030)	(0.042)	(0.021)
<5 yrs of experience	0.015	0.012	0.015	0.014**	-0.067	0.065	0.086**	0.032
	(0.015)	(0.011)	(0.011)	(0.007)	(0.058)	(0.056)	(0.033)	(0.028)
Yrs of experience	-0.001	0.000	0.001	-0.000	-0.007**	0.001	-0.001	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.004)	(0.002)
Observations	15,314	17,048	16,952	49,314	15,200	16,946	16,866	49,012
Adjusted R-squared	0.003	0.033	-0.001	0.012	0.509	0.515	0.498	0.507

Notes: *,**,*** denotes statistically significance at the 10, 5 and 1 percent-level, respectively.

Appendix Table 6: Classroom and Teacher Value-Added Estimates: Same Sample of Teachers, Control Schools

Panel A: Leblango Reading	Classroom Effects (3)	Teacher Effects (4)
Corrected SD of effects	0.24	0.19
	[0.191,0.279]	[0.133,0.239]
Observations (pupil-by-year)	11,430	11,430
Pupils	7,468	7,468
Teachers	146	146
Classrooms	347	347
Schools	40	40
Pupils per classroom/teacher	45	78
Panel B: English Reading	Classroom Effects (3)	Teacher Effects (4)
Corrected SD of effects	0.26	0.21
	[0.208,0.312]	[0.168,0.249]
Observations (pupil-by-year)	6,116	6,116
Pupils	4,360	4,360
Teachers	99	99
Classrooms	211	211
Schools	40	40
Pupils per classroom/teacher	38	55

Notes: The estimates are based on the Two-teacher and Longitudinal Samples conditioning on teachers being in both samples. 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 school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table 7: Robustness Estimates of Teacher Value-Added: Restricting to Classes with Minimum of 10 or 15 Students, School effects purged, Control Schools

	Minimum of 10 Students		Minimum of 15 Students	
Panel A: Leblango Reading	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Corrected SD of effects	0.26	0.18	0.25	0.25
	[0.278,0.375]	[0.172,0.291]	[0.294,0.352]	[0.294,0.352]
Observations (pupil-by-year)	17184	11231	16449	10796
Pupils	8750	7394	8577	7257
Teachers	329	145	296	137
Classrooms	504	318	440	280
Schools	42	40	41	39
Pupils per classroom/teacher	44	77	45	79
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal
	Minimum of 10 Students		Minimum of 15 Students	
Panel B: English Reading	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Corrected SD of effects	0.28	0.16	0.28	0.17
	[0.263,0.400]	[0.139,0.238]	[0.267,0.389]	[0.139,0.274]
Observations (pupil-by-year)	10596	6000	10063	5708
Pupils	5613	4294	5506	4152
Teachers	256	95	227	96
Classrooms	347	194	300	168
Schools	42	40	41	39
Pupils per classroom/teacher	38	56	39	57
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal

Notes: The estimates are based on the Two-teacher and Longitudinal Samples conditioning on class sizes being larger than 10 or 15 students. 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 school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table 8: Robustness Estimates of Teacher Value-Added: Dropping or Alternative Imputation of Test Scores, School Effects Purged, Control Schools

Panel A: Leblango Reading	Dropping student-year observations with missing baseline or prior-year scores		Omitting grade-one student-year observations	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Corrected SD of effects	0.34	0.20	0.32	0.19
	[0.332,0.492]	[0.196,0.348]	[0.277,0.368]	[0.139,0.195]
Observations (pupil-by-year)	13920	9177	13,946	8,449
Pupils	7868	6408	6,324	5,323
Teachers	353	144	325	132
Classrooms	540	331	454	259
Schools	42	40	42	40
Pupils per classroom/teacher	39.69899368	63.7291667	43	63
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal

Panel B: English Reading	Dropping student-year observations with missing baseline or prior-year scores		Omittng grade-two student-year observations	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
	(1)	(2)	(3)	(4)
Corrected SD of effects	0.35	0.22	0.26	0.20
	[0.298,0.435]	[0.141,0.351]	[0.321,0.378]	[0.124,0.321]
Observations (pupil-by-year)	7190	3908	6,508	3,209
Pupils	4102	2938	4,623	2,885
Teachers	268	103	188	62
Classrooms	359	194	245	119
Schools	42	40	42	36
Pupils per classroom/teacher	29	37	35	48
Sample	Two-Teacher	Longitudinal	Two-Teacher	Longitudinal

Notes: The estimates are based on the Two-teacher and Longitudinal Samples. 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 school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table 9: Robustness Estimates of Teacher Value-Added: Purging School-Year Effects, Control Schools

Panel A: Leblango Reading	Purging School-Year Effects instead of School Effects	
	Classroom Effects	Teacher Effects
	(3)	(4)
Corrected SD of effects	0.19	0.18
	[0.204,0.292]	[0.189,0.350]
Observations (pupil-by-year)	15,721	10,104
Pupils	8,454	6,974
Teachers	356	143
Classrooms	540	327
Schools	42	42
Pupils per classroom/teacher	39	70
Sample	Two-Teacher	Longitudinal

Panel B: English Reading	Purging School-Year Effects instead of School Effects	
	Classroom Effects	Teacher Effects
	(3)	(4)
Corrected SD of effects	0.20	0.21
	[0.191,0.276]	[0.182,0.338]
Observations (pupil-by-year)	10334	5723
Pupils	5557	4141
Teachers	281	94
Classrooms	377	202
Schools	42	42
Pupils per classroom/teacher	36	53
Sample	Two-Teacher	Longitudinal

Notes: The estimates are based on the Two-teacher and Longitudinal Sample. Students with missing baseline scores or characteristics are dropped. 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 school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table 8: Robustness Heterogeneity of Value-Added by NULP Study Arm, 2017 Data Omitted and only Treated Teachers

	2017 Data Omitted						Only Treated Teachers					
	Classroom Effects			Teacher Effects			Classroom Effects			Teacher Effects		
	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
Panel A: Leblango Reading	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Corrected SD of effects	0.22	0.33	0.40	0.10	0.15	0.23	0.24	0.35	0.42	0.18	0.24	0.30
	[0.212,0.340]	[0.330,0.439]	[0.358,0.456]	[0.085,0.200]	[0.159,0.236]	[0.188,0.299]	[0.231,0.314]	[0.350,0.451]	[0.391,0.476]	[0.172,0.264]	[0.230,0.389]	[0.274,0.368]
Observations (pupil-by-year)	13,785	15,206	14,887	8,795	10,050	10,470	13,126	14,752	14,748	10,310	12,315	12,665
Pupils	8,579	9,441	8,990	6,975	8,000	7,732	7,654	8,753	8,428	6,855	8,340	7,901
Teachers	277	292	262	101	114	112	214	225	207	126	148	146
Classrooms	423	461	438	253	289	293	395	436	425	306	359	362
Schools	42	44	42	41	44	42	42	44	42	40	44	41
Pupils per classroom/teacher	33	33	34	34	34	36	33	34	35	33	34	35
Sample	Two-Teacher			Longitudinal								
Panel B: English Reading	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Corrected SD of effects	0.24	0.32	0.34	0.11	0.17	0.23	0.26	0.31	0.34	0.20	0.24	0.24
	[0.215,0.356]	[0.314,0.462]	[0.255,0.393]	[0.115,0.144]	[0.164,0.309]	[0.166,0.296]	[0.220,0.342]	[0.306,0.421]	[0.264,0.395]	[0.132,0.229]	[0.222,0.399]	[0.181,0.315]
Observations (pupil-by-year)	7,055	7,758	7,803	3,491	3,793	4,280	7,412	8,002	8,088	5,608	6,315	6,564
Pupils	5,036	5,465	5,469	3,264	3,457	3,726	4,949	5,282	5,215	4,150	4,660	4,760
Teachers	194	202	187	45	44	56	157	158	148	86	84	94
Classrooms	245	263	246	118	127	128	252	263	247	189	204	199
Schools	42	44	42	39	38	37	42	44	42	40	43	41

Pupils per classroom/teacher	29	29	32	24	24	31	29	30	33	29	29	32

Notes: The sample includes the Two-teacher Sample for classroom effects and the Longitudinal Sample for teacher effects. All estimates are purged of school effects by subtracting off the school mean. 95% confidence intervals for the SD of the classroom effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.

Online Appendix

Online Appendix Table 1: Correlation between Student Attrition and Student Characteristics

Student characteristics	Control (1)	Full-cost (2)	Reduced-cost (3)	All (4)
Female (1=Yes)	-0.247***	-0.244***	-0.228***	-0.248***
	(0.009)	(0.015)	(0.006)	(0.009)
Female × Full-cost				0.021*
				(0.011)
Female × Reduced-cost				0.003
				(0.018)
Age	-0.019***	-0.015***	-0.014***	-0.017***
	(0.003)	(0.002)	(0.002)	(0.002)
Age × Full-cost				0.003
				(0.003)
Age × Reduced-cost				-0.002
				(0.003)
Full-cost program				-0.064**
				(0.027)
Reduced-cost program				-0.017
				(0.026)
Observations	23,676	25,691	24,689	74,056
Adjusted R-squared	0.149	0.149	0.130	0.142

Attrittion defined within years (ie. present at baseline but missing at endline within the same year)

Online Appendix Table 2: Correlation between Student Attrition and Teacher Characteristics

Teacher characteristics	Control (1)	Full-cost (2)	Reduced-cost (3)	All (4)
Female (1=Yes)	-0.004	0.003	-0.005	-0.007**
	(0.003)	(0.005)	(0.005)	(0.003)
Female × Full-cost				0.004
				(0.006)
Female × Reduced-cost				0.012**
				(0.006)
Age	-0.000	0.002*	0.000	0.000
	(0.000)	(0.001)	(0.001)	(0.000)
Age × Full-cost				-0.000
				(0.001)
Age × Reduced-cost				0.001
				(0.001)
> Bachelor (1=Yes)	-0.001	-0.001	0.002	-0.007*
	(0.003)	(0.007)	(0.006)	(0.004)
> Bachelor (1=Yes) × Full-cost				0.011
				(0.007)
> Bachelor (1=Yes) × Reduced-cost				0.007
				(0.008)
< 5 yrs of experience (1=Yes)	0.002	0.012	-0.011	0.004
	(0.002)	(0.013)	(0.006)	(0.004)
< 5 yrs of experience (1=Yes) × Full-cost				-0.015**
				(0.007)
< 5 yrs of experience (1=Yes) × Reduced-cost				0.008
				(0.013)
Experience (years)	0.000	-0.002*	0.000	-0.000
	(0.000)	(0.001)	(0.001)	(0.000)
Experience × Full-cost				0.001
				(0.001)
Experience × Reduced-cost				-0.001
				(0.001)
Full-cost program				0.002
				(0.022)
Reduced-cost program				-0.035
				(0.025)
Observations	15,320	17,072	16,993	49,385
Adjusted R-squared	0.001	0.010	0.003	0.008

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Online Appendix Table 3: Correlation between Teacher Attrition and Teacher Characteristics

Teacher characteristics	Control (1)	Full-cost (2)	Reduced-cost (3)	All (4)
Female (1=Yes)	-0.236***	-0.137***	0.156***	-0.236***
	(0.064)	(0.048)	(0.055)	(0.063)
Female × Full-cost				0.080
				(0.084)
Female × Reduced-cost				0.099
				(0.080)
Age	-0.002	-0.007	0.009	-0.002
	(0.006)	(0.007)	(0.007)	(0.006)
Age × Full-cost				0.011
				(0.009)
Age × Reduced-cost				-0.005
				(0.009)
> Bachelor (1=Yes)	0.039	-0.077	0.043	0.039
	(0.083)	(0.066)	(0.077)	(0.082)
> Bachelor (1=Yes) × Full-cost				0.004
				(0.112)
> Bachelor (1=Yes) × Reduced-cost				-0.117
				(0.105)
< 5 yrs of experience (1=Yes)	0.148	-0.061	-0.054	0.148
	(0.114)	(0.128)	(0.137)	(0.113)
< 5 yrs of experience (1=Yes) × Full-cost				-0.202
				(0.177)
< 5 yrs of experience (1=Yes) × Reduced-cost				-0.209
				(0.170)
Experience (years)	-0.001	-0.006	-0.012	-0.001
	(0.008)	(0.007)	(0.008)	(0.008)
Experience × Full-cost				-0.011
				(0.011)
Experience × Reduced-cost				-0.006
				(0.010)
Full-cost program				-0.338
				(0.271)
Reduced-cost program				0.245
				(0.270)
Observations				
Adjusted R-squared	266	291	272	829

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1