Learning More about Teachers: Estimating Teacher Value-Added and Treatment Effects on Teacher Value-Added in Northern Uganda

Julie Buhl-Wiggers, Jason T. Kerwin, Jeffrey Smith, and Rebecca Thornton*

May 3, 2022

Abstract

This paper uses longitudinal data from a school-based RCT in northern Uganda to estimate teacher value-added. We first provide lower bounds on the variation in teacher effectiveness – 0.23 SDs in local-language reading and 0.19 in English reading – the first estimates from sub-Saharan Africa. Second, comparing our estimates under years of random assignment of students to classrooms with years under business-as-usual assignment, we find no evidence of non-random sorting of students to teachers. Third, we measure the causal effects of providing high-impact teacher training and support and find the variation in teacher effectiveness increases by 52% in local language reading, likely by improving already-effective teachers the most. Fourth, we find that observed teacher characteristics are weakly correlated with teacher effectiveness and the gains in quality caused by the training.

JEL Codes: I2, O1

Keywords: Teachers, RCT, Africa, Value-Added

*Buhl-Wiggers: Department of Economics, Copenhagen Business School (jubu.eco@cbs.dk); Kerwin: Department of Applied Economics, University of Minnesota and J-PAL (jkerwin@umn.edu); Smith: Department of Economics, University of Wisconsin (econjeff@ssc.wisc.edu) Thornton: Department of Economics, University of Illinois (rebeccat@illinois.edu). We thank Laura Schechter and Chao Fu and seminar audiences at the University of Minnesota, CSAE, RISE, the University of Wisconsin, SOLE, the University of Alabama, the University of Missouri, the University of Houston, Baylor University, UCSC, Purdue, UCR, Stellenbosch University, the International Population Conference, NYU Steinhardt, Georgetown, UT Austin, Northeastern University, the "Mike and Scott" online Economics of Education Seminar, and the NBER Summer Institute 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. The usual disclaimer applies.

1 Introduction

Extensive evidence shows that the most important predictor of student learning is the quality of the student's teacher. Studies from the United States show that having a more effective teacher can substantially affect student learning and long-run outcomes (e.g. Rivkin, Hanushek, and Kain 2005; Chetty, Friedman, and Rockoff 2014; Chetty et al. 2011); recent studies echoe this in Latin America and Asia (e.g. Araujo et al. 2016; Bau and Das 2020; Azam and Kingdon 2015). One implication of these findings is that moving less effective teachers to the level of the best would increase student learning and decrease inequality in education. Although a large literature has provided evidence on the return to specific educational inputs, little is known about how different interventions affect the variation in teacher quality. Using five years of panel data from a randomized evaluation of the Northern Uganda Literacy Project (NULP), we provide the first causal estimates of an effective teacher-focused intervention on the variation in teacher effectiveness. By focusing on an extreme setting of poverty in sub-Saharan Africa, we in addition bring forth new evidence on teacher effectiveness in developing countries.[1]

This paper contributes four new insights. First, we show that there is substantial variation in teacher effectiveness in Uganda. We find that a one standard deviation increase in teacher value-added improves local language reading test scores by 0.23 standard deviations and improves English reading by 0.19 standard deviations; teacher value-added is strongly correlated across subjects, with a correlation coefficient of 0.73 between English and local language reading. Our estimate of the standard deviation of teacher effects on local language reading is over twice as large as the Chetty, Friedman, and Rockoff (2014) estimate for the effect of American primary-school teachers on native-language reading scores, as well as larger than 0.13 SDs, the average across nine studies in the US, reviewed by Hanushek and Rivkin (2012). Our estimates are also larger in native reading than those found among kindergarteners in Ecuador (0.09 SDs; Araujo et al. 2016). In Pakistan, estimates for reading are 0.06 SDs in Urdu, and 0.17 SDs in English (Bau and Das 2020).[2]

Second, little is known about the dynamics of assignment of students to teachers of particular quality within a classroom in Africa, or in lower-income settings more generally. We first provide survey evidence from head teachers in our study, who report how students

[1] 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 (2019), for Uganda, Blackmon (2017), for Tanzania, and Muñoz-Chereau and Thomas (2016), for Chile. Oketch, Rolleston, and Rossiter (2021) estimate classroom value added in Ethiopia, but do not isolate the independent effects of teachers.

[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. Second, they focus on teachers in secondary, rather than primary schools.

are assigned to teachers. Then, utilizing the fact that we have the same teachers teaching in both random and business-as-usual years, we shed light on the degree of bias arising from non-random sorting of students to teachers. Our estimates of classroom value-added are nearly the same when we compare business-as-usual years to random assignment years, consistent with evidence from the US (e.g. Kane and Staiger 2008). The finding of limited systematic sorting helps ensure that the our estimates of teacher value-added are not driven by students systematically sorting into classrooms and that we can interpret our results as causal—i.e. as answers to the question: "how does being assigned to a particular teacher affect student achievement?" (Rothstein 2010; Goldhaber and Chaplin 2015).

Third, existing literature on value-added provides information about the distribution of teacher quality but is not necessarily informative about policies aimed at improving teacher quality. Using the fact that the NULP teacher-focused intervention was randomly assigned across schools, we compare the variation in teacher value-added across treatment arms to estimate the first causal obtain the effects of teacher training on the variance of teacher effectiveness.

Schools were randomized to three study arms: 1) a full-cost version of the literacy program, 2) a reduced-cost version of the literacy program, and 3) a control. In schools assigned to the full-cost program, the NULP was delivered directly to teachers. In reduced-cost schools, teacher training and support was conducted using a cascade model in collaboration with government tutors. Schools in the control group did not receive the NULP. Both NULP versions 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.[3]

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 52 percent in full-cost program schools; both effects are statistically significant at the five percent-level.

Finally, we examine which – if any – teacher characteristics are correlated with our estimated measures of classroom and teacher value-added, and characteristics that predict treatment effect heterogeneity on the distribution of value-added. The literature that attempts to characterize high-quality teachers by correlating teacher value-added with observed

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

characteristics typically finds that the first years of experience are important, with few other successful predictors (Azam and Kingdon 2015; Slater, Davies, and Burgess 2012; Araujo et al. 2016; Bau and Das 2020). Our findings are generally consistent with those prior studies. We find that having a bachelor's degree is negatively associated with classroom and teacher value-added, while gender and years of experience appear uncorrelated. Similar to prior literature, we can explain very little of the variation in effectiveness using teacher characteristics (less than two percent in local language reading).

Correlating teacher characteristics with our estimates of the treatment effects of the NULP on value-added, we find some suggestive evidence of larger effects of the NULP among teachers with fewer years of experience. We also make some headway in understanding which teachers benefit most from the program. Testing for rank preservation we cannot reject that the program had rank-preserving impacts on teacher effectiveness suggesting that the NULP achieved its impacts by improving teaching primarily among the most-effective teachers.

In addition to our main results, we present a number of sensitivity analyses that show our results are robust to alternate choices about how we construct our sample and analyze the data.

Measures of the distribution of teacher value-added are typically used to gauge the scope by which teachers are able to affect student learning. At the extreme, if the variance is close to zero, a child's teacher is perfectly substitutable with another, without having any (positive or negative) impact on learning. A wider distribution in value-added suggests to many in this literature area a larger scope for policy makers to substitute a bad teacher for a better one. This logic makes sense if the best teachers are teaching at their production frontier.

In rural Africa or other low-resource settings where the best teachers still have room for improvement, using the distribution of teacher value-added as this type of metric may not be as useful. The fact that the NULP intervention massively increased student learning and resulted in a wider distribution of value-added–by moving the teachers who were already at the upper end of the teaching effectiveness distribution–suggests precisely that we should interpret these measures of value-added with caution. We hope that more studies will examine what can help support teachers. Just as we have moved past simply understanding how the average student is affected by education interventions (see Buhl-Wiggers et al. 2022), we hope future research will examine how to support teachers at all places on the distribution of teaching quality.

xx

2 Setting and Intervention

This section describes the setting of the study in northern Uganda, the specific teacher intervention we evaluate, and the data we use to do so.

2.1 Setting: Education in Uganda

Primary education in Uganda consists of seven years of schooling, 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. While the country's net enrollment rate is now above 90 percent, only about 60 percent of students transition from primary to secondary school (Deininger 2003; World Bank 2020). Uganda also faces major learning challenges in its schools. Bold et al. (2017) find that the vast majority (94 percent) of children in government primary schools could not read a simple paragraph. Among students in grade seven, 20 percent are unable to read and understand a short story (Uwezo 2016).

In Uganda, there are 11 different languages of instruction and in 2007, the government mandated local language instruction in the lower primary grades (one to three). 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, and the education system does not have 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 2014). Teachers in Uganda receive in-service training referred to as Continuous Professional Development (CPD) which is intended to update competencies 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 head teachers. 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).

Teachers in Uganda, as in sub-Saharan Africa more generally, face severe constraints on their ability to teach effectively: they are undertrained, lack quality materials and methods for teaching, face crowded classrooms, and work in schools with nonexistent systems for tracking pupil performance and insufficient school supervision. Bold et al. (2017) 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.

2.2 NULP: Intervention

The Northern Uganda Literacy Project (NULP) is an early-grade mother-tongue literacy program developed in response to the educational challenges facing northern Uganda. Of the four regions in Uganda, Northern Uganda is the poorest, with a history of marginalization. The region contains only 20 percent of the population, yet almost half of the poorest 20 percent of Ugandans live in northern Uganda (Ministry of Finance 2003). The area experienced decades of civil war leading to millions of internally displaced people and severe infrastructure shortages. More recently, the area has experienced large flows of refugees from South Sudan. This historical context has resulted in an overstretched and poorly-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 a week-long residential teacher training three times a year and monthly classroom support visits to give feedback to teachers. The program's approach involves training teachers how to be more engaged with students and move 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.[4]

The full-cost version of the 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 imple-

[4] 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.

mented in some schools, following a "cascade" or "training-of-trainers" delivery model led by Ministry of Education CCTs rather than Mango Tree staff; teachers in these schools also received fewer support visits.[5]

The NULP was introduced to different grades during our study (Appendix Table A1, 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 the program.[6] 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.

2.3 NULP: Sample of Schools, Students and Teachers

2.3.1 Schools

Schools were sampled 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.[7] 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.[8]

2.3.2 Students

We follow four cohorts of first-grade children who entered the study schools in 2013, 2014, 2015, and 2016, comprising a total of 29,787 students (Table 1, Panel A). Appendix Table A2 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 near the end of 2014 (Cohort 1 endline sample). In 2014, 100 first-grade students were randomly selected

[5] 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

[6] 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 training or support were delivered. This intervention was much less intensive than the earlier years

[7] The other eligibility criteria for 2013 were desks and lockable cabinets for each grade 1 class, a student-to-teacher ratio in grade 1 to grade 3 of no more than 135 in 2012, being located less than 20 km from the main district school coordinating offices, being accessible by road year round, having a head teacher regarded as "engaged", and not having previously received support from Mango Tree.

[8] The other eligibility criteria for 2014 were having desks and blackboards in grade 1 to grade 3 classrooms and having a student-to-teacher ratio of no more than 150 students during the 2013 school year in grade 1 to grade 3.

from each of the 128 schools—sampled either at baseline or endline (Cohort 2).[9] In 2015, 30 first-grade students (Cohort 3) were randomly selected from each school at endline. Lastly, in 2016, 60 first-grade students (Cohort 4) were randomly selected from each school and 30 additional second-grade students were added to Cohort 3 at endline. For each cohort, the sample of students was stratified by gender and classroom.

2.3.3 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 teachers per grade. Table 1 shows that in our sample, on average, there are approximately two teachers per grade; this varies across year and school.

[Table 1 about here.]

2.4 Randomization

2.4.1 Random Assignment of NULP to Schools

Schools in the study were assigned to one of three study arms: 1) full-cost NULP, 2) reduced-cost NULP, and 3) control. Schools were grouped into stratification cells of three schools each.[10] Each stratification cell contained 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 12 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.

2.4.2 Random Assignment of Students to Teachers

The assignment of students to classrooms in Uganda is specific to each school and depends on the approach used by the school's head teacher. In three of the five years of the study

[9] 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, 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.

[10] 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.

(2013, 2016 and 2017), we explicitly instructed head teachers to randomly assign students to classrooms (Appendix Table A1, Panel B).[11] In 2014 and 2015, head teachers were not given any guidance on how to assign students to classrooms. Within a school, head teachers have discretion to assign teachers to specific grades.

Our analysis in this paper does not account for sorting of teachers to particular schools or grades.

2.5 Data

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

2.5.1 Learning Outcomes: Student Reading Test Scores

Our student learning outcomes consist of test scores in local language reading and English reading. Test administration varied somewhat by subject, year, and cohort, summarized in Appendix Table A1, 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.[12]

The tests involved the Early Grade Reading Assessment (EGRA), an internationally recognized assessment of early literacy skills (Dubeck and Gove 2015; RTI 2009; Piper 2010; Gove and Wetterberg 2011). We use two different validated versions of the test—English and Leblango.[13] For each language, we construct indices by first standardizing the separate test components against the control group for each student-year-grade observation, and second, constructing a principal component score index for the entire assessment using the factor loadings from the control group in grade 3 in 2016. These indices for local language and English reading are then standardized against the control group separately for each year and grade.

[11] 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.

[12] This results in Cohort 4 students only being assessed in one year, when they were in grade one in 2016.

[13] 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.

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 assessments were conducted beginning in grade two; first-grade students were administered an oral English test.

2.5.2 Student Characteristics

Our student-level analyses control for age and gender. In addition to controlling for prior year reading scores (described below), we also include a control for pre-intervention math ability based on 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.[14]

2.5.3 Teacher Characteristics

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

3 Theoretical Considerations

In this section, we briefly describe the canonical teacher value-added model and implications of our low-income setting. We also outline predictions of the effects of an educational intervention on the distribution of teacher quality.

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 an estimate of the increase in learning associated with a specific classroom or teacher.

The variance of teacher value-added represents the scope for improving student performance– by either moving teachers up the distribution or removing teachers at the lower end of the distribution. Assuming normality, a variance in teacher value-added of 0.20 SDs suggests

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

that having a teacher at the 75th percentile (as compared to the 25th percentile) of the quality distribution, increases student test scores by 0.27 SDs; a variance of 0.10 SDs suggests that such a shift would only increase student test scores by 0.14 SDs. The usefulness of this interpretation hinges on two possible policy actions: either it is possible to move the worst-performing teachers to the level of the best, or it is possible to identify and replace the worst performing teachers with more-effective teachers.

In the existing value-added literature, estimates of teacher value-added represent how close or far teachers are from those who are performing at their best (Sass et al. 2012). Implicit is the assumption that lower variance teacher value-added implies that more teachers are teaching at their highest capacity. In settings such as the United States, it is possible that high-performing teachers are working at the frontier of the set of education production possibilities. In lower-income settings however, a lower estimated variance of teacher value-added may simply imply that most teachers are performing poorly, with little gains to student learning. Schooling outcomes from sub-Saharan Africa have been found to be among the worst on the planet; our data from rural northern Ugandan, provide estimates among some of the poorest schools, teachers, and students, that have been measured.

How would we expect an education intervention focused on teachers to affect the distribution of teacher quality? To the extent that the intervention improved all teachers equally, we may not expect any change in the distribution of teacher value-added. A decrease in the variance of teacher value-added suggests that low-performing teachers experienced a greater benefit as a result of the training, for example if these teachers had more to gain than higher-performing teachers. On the other hand, if high-ability teachers benefit more than lower ability teachers ("skills begets skills"), the variance of the teacher value-added would increase as a result of training.

While previous research is able to estimate the scope for test score improvements if the worst performing teachers were to be hypothetically moved 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. The NULP program was highly effective, resulting in massive effects on student learning. Comparing across treatment arms, Buhl-Wiggers et al. (2018) find that after three years of exposure to the program, the full-cost program increased local-language reading scores increasing by 1.35 SDs and the reduced-cost program increased reading by 0.78 SDs.

In our setting, rather than estimating the hypothetical effect of moving teachers from high to low value-added, we test what actually happens to the variation in teacher value-added with increases in average teacher quality. Because the NULP was randomly assigned across schools, we provide the first causal estimates of the effects of teacher training on the variance

of teacher effectiveness. To understand the types of teachers who experience gains from the NULP, we present interactions with teacher characteristics, as well as conduct formal tests for rank preservation to understand which teachers are likely affected by the program.

4 Empirical Approach

In this section we describe our main analytical samples and strategies to estimate classroom and teacher value-added, compare estimates under random and business-as-usual student to teacher assignment, measure the causal effects of the NULP on value-added, and understand which teachers are highest quality and improve most under the NULP intervention.

4.1 Analytical Samples

4.1.1 Annual Student Learning Gains

Our analytical strategy involves measuring the average gain in student learning attributable to a teacher in a given school year. Appendix Table A3 provides a detailed description of the tests used to estimate value-added for each subject, grade, and year of the study. For each student, we need an endline test score for any given year; we drop student-year observations in which a student is missing an endline test in local language or English reading. This results in 58,775 student-year observations and 27,943 unique students (Table 1, Panels A and B).

Next, for every student-year observation with an endline test, we identify 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.[15]

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 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 (See Appendix Table A3).

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 assessment. In that case, we

[15] This is motivated by the fact that 1) we only have baseline tests 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 tests, or only impute scores that are missing.

impute students' missing prior test score as zero.[16] 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 below) to address missing prior scores.

4.1.2 Classroom and Teacher Effects Samples

To estimate classroom and teacher value-added we match students to specific teachers using classroom registers and student reports. Across 58,775 total student-year observations for which we have at least one endline test score, we are able to match almost all to a teacher (99 percent).[17] The most common reasons for not being able to match students to teachers include missing or misreported teacher names.[18] To limit estimation error due to sampling variation, we drop student-year observations with fewer than five students per teacher in a given year. This removes 2,188 observations, corresponding to 3.8 percent of the overall sample, and bringing us down to 56,035 student-year observations (Table 1, Panel C).[19]

To estimate classroom and teacher effects, 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 that 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 D shows the number of schools and teachers meeting this criterion, forming our Two-Teacher Sample: 56,035 student-year observations (27,609 students) and 1,763 classrooms (1,096 teachers).

To separate teacher effects from classroom effects, we need to observe a teacher over multiple years. We observe 475 (34 percent) of teachers teaching at least two years (38

[16] 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).

[17] 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).

[18] 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.

[19] The rate of observations with fewer than 5 students per teacher 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; the p-value from an F-test testing the equality across treatment arms is 0.12).

percent in the full-cost treatment, 34 percent in reduced-cost treatment, and 31 percent in the control).[20] This is our Longitudinal Sample and includes 1,138 classrooms (475 teachers) and 38,082 student-year observations (24,218 students). See Table 1, Panel E.

4.1.3 Teacher Characteristics Sample

Table 1, Panels D and E present the number of teachers for whom we have teacher characteristic data. Of the 1,096 teachers in our *Classroom Effects Sample*, we have teacher characteristics for 871 (79 percent); 81 percent in the full-cost program, 80 percent in the reduced-cost program, and 77 percent in the control group. Of the 475 teachers in the *Teacher Effects Sample*, we have characteristics for 435, or 91 percent with similar rates across the study arms.

4.2 Balance and Attrition

4.2.1 Balance across NULP Treatment Arms

Appendix Table A4 presents descriptive statistics for students and teachers in each of our analytical samples, separated by study arm. Half of students are female (recall that the sample is stratified by gender), and students are on average almost nine years old (Panel A). On average, teachers are around 40 years old, 48 percent female, with 14 years of education and 14 years of experience (Panel B).

Schools are generally balanced across study arms in terms of student characteristics—age and gender—and teacher characteristics.

4.2.2 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 pre-intervention student characteristics, which gives us an indication of whether certain students are matched to certain teachers. Appendix Table A8 presents regressions of student pre-intervention test scores on teacher characteristics. While there are a few statistically significant coefficients, the majority are small and insignificant.

As a second check for balance across randomly assigned students to teachers, we test the difference in student prior test scores between classes within schools and grade levels for each year, which indicates the degree of sorting similar students into the same classes (Horvath 2015). Appendix Figure A1 presents a distribution of p-values from regressing baseline test

[20] Conducting a test of equality of the school-level rates across treatment arms gives a p-value of 0.78

scores on classroom dummies within each year, school and grade-level. We find that around 4 percent of the schools had classrooms with statistically significant (at the 5-percent level) baseline differences across classroom streams in the random assignment years (2013, 2016 and 2017), which is what we would expect by random chance.

4.2.3 Student and Teacher Attrition

Student attrition from the study could be due to dropping out, transferring to another school, or being absent for an assessment. The extent to which certain types of students attrit—either overall or differentially by study arm—could affect the external and internal validity of our analysis. Appendix Table A5 presents the correlation between student characteristics and student attrition. In general, attritors 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.[21]

Teacher attrition is an important issue, given that our Teacher Effects Sample requires observing a teacher over at least two years. Appendix Table A6 presents the correlation between teacher characteristics and teacher attrition.[22] Female teachers are less likely to attrit, however, this does not vary between study arms.[23] Appendix Table A7 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.

4.3 Estimation Strategy

This section describes our empirical approach to estimating classroom and teacher value-added and the causal effects of the NULP.

4.3.1 Classroom and Teacher Effects

We begin by estimating classroom effects using the following "lagged-score" value-added model, separately for local language reading and English reading:[24]

[21] We define student attrition as a missing student-year observation of test scores and examine 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.

[22] Teacher attrition is defined as teachers only being in our two-teacher sample. Thus, we observe them once and then they drop out of our sample.

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

[24] In a simulation exercise, Guarino et al. (2015) find, that the "lagged-score" model performs best in most scenarios. Our results are fairly similar to if we using use a "gain-score" model, in which we do not control for lagged test scores and instead replace the left-hand-side of Equation (1) with $\Delta Y_{icgs,t} = Y_{icgs,t} - Y_{icgs,t-1}$ (see Appendix Table A9, Columns 3 and 4).

$$
\begin{aligned}
Y_{icgs,t} =& \beta_1 Y_{icgs,t-1} + \beta_2 Z_{icgs,t-1} + \beta_3 X_{icgs,t} + \lambda_{cgs,t} + \zeta_g + \beta_6 D_{icgs,t} + \beta_7 ST_{icgs,t} + \\
& \beta_4 Y_{icgs,t-1}\zeta_g + \beta_5 Z_{icgs,t-1}\zeta_g + \epsilon_{icgs,t}
\end{aligned} \tag{1}
$$

where $Y_{icgs,t}$ is the endline test score (Leblango or English) for child i in classroom c, in grade g, in school s, in year t. $Y_{icgs,t-1}$ is the student's prior test score for the test of interest.[25] $Z_{icgs,t-1}$ is a vector of prior scores for the other reading assessment and math. Both of these capture previous family, school and individual factors as well as genetic endowments. $X_{icgs,t}$ is a vector of individual characteristics, specifically gender and age. The $\lambda_{cgs,t}$ are classroom fixed effects; year fixed effects are implicit in 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 use indicators for whether prior test scores, age or gender are missing $D_{icgs,t}$. Moreover, we include an indicator for the sample type $ST_{icgs,t}$, which is equal to one if the child was sampled at endline and zero for students in the baseline sample. 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_4 and β_5. To estimate a full set of classroom effects, we omit the constant term from the regression.

$\lambda_{cgs,t}$ is the effect of being in a specific classroom, and thus $\hat{\lambda}_{cgs,t}$ is an estimate of the increase in learning attributable to a specific classroom and teacher in year t. We use use all possible observations meaning those observations that can be matched to a teacher (58,223 student-year obs) to estimate $\lambda_{cgs,t}$. After estimating $\lambda_{cgs,t}$ we restrict our sample to either the Classroom Effects Sample or the Teacher Effects sample as described above.

The estimated classroom effects from Equation (1) contain both a permanent teacher component as well as a transitory classroom component that captures things like disturbances during testing or peer dynamics during a particular year. 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. We estimate teacher effects using the classroom effects with the following equation:

$$
\hat{\lambda}_{cgs,t} = \hat{\delta}_{cgs} + \omega_{cgs,t} \tag{2}
$$

where, $\hat{\delta}_{cgs}$ is a vector of teacher indicators and can be interpreted as the "permanent" teacher component. 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. The identifying

[25] For grade 1 these are all zero. For grades 2 and above this is prior end-of-year test scores.

assumption is that $\omega_{cgs,t}$ is not serially correlated across years. If this assumption fails, $\omega_{cgs,t}$ could contain "real" teacher quality fluctuations, and $\hat{\delta}_{cgs}$, would be biased toward zero.

The variation in teacher effects ($var(\hat{\delta}_{cgs})$) can then be interpreted as the scope to which it matters which teacher a student is assigned to; small variance means that teachers are very similar and thus it doesn't matter which teacher a student gets; on the contrary a large variance means that teachers are very different and thus it matters a great deal which teacher a student is assigned to.

In interpreting estimates of λ_{cgst} and δ_{cgs}, three issues arise: First, there may be school effects or school-level shocks that co-vary with true classroom and teacher effects due to factors such as school management or school quality. Second, the estimated classroom and teacher effects are the sum of the true classroom and teacher effects and the estimation error that arises from the fact that we have relatively small samples of students per classroom and 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 and teacher effects. Third, there may be individual student effects that co-vary with true classroom effects due to sorting of students to teachers based unobserved characteristics. We address each of these three issues in turn.

4.3.2 Purging School Effects From Classroom and Teacher Effect Estimates

When estimating Equation (1) we use both within- and between-school variation. This means that the estimate $\hat{\lambda}_{cgs,t}$, 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 evident variation in our estimated classroom effects likely results from 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):

$$\hat{\gamma}_{cgst} = \hat{\lambda}_{cgst} - \hat{\lambda}_{s} \tag{3}$$

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.

In the same manner we de-mean our estimated teacher effects ($\hat{\delta}_{cgs}$) by the school average to purge any school effects:

$$\hat{\zeta}_{cgs} = \hat{\delta}_{cgs} - \hat{\delta}_s \tag{4}$$

4.3.3 Removing Sampling Variation

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 finite samples of students. The smaller the number of students, the more likely that the estimated effect on learning of a given classroom will be very large or small 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).[26] 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:[27]

$$\hat{V}_{corrected}(\hat{\gamma}_{cgs,t}) = V(\hat{\gamma}_{cgs,t}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{[(\sum_{c=1}^{C_s} N_{cs}) - N_{cs}]}{N_{cs}(\sum_{c=1}^{C_s} N_{cs})}\hat{\sigma}^2\right\} \tag{5}$$

where $\hat{\sigma}^2$ is the variance of the estimated residuals, $\hat{\epsilon}_{icgs,t}$, 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. $\hat{V}_{corrected}(\hat{\gamma}_{cgs,t})$ is our measure of interest when discussing the distribution of classroom effects.

We correct the variance of the teacher effects for sampling variation using the following adjusted form of Equation (5):

$$\hat{V}_{corrected}(\hat{\zeta}_{cgs}) = V(\hat{\zeta}_{cgs}) - \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{[(\sum_{t=1}^{T_s} N_{ts}) - N_{ts}]}{N_{ts}(\sum_{t=1}^{T_s} N_{ts})}\hat{\sigma}^2\right\} \tag{6}$$

where $\hat{\sigma}^2$ is the variance of the residuals, $\hat{\epsilon}_{icgst}$, from Equation (1). T is the overall number of teachers in the sample, and N_{ts} is the number of students taught by teacher t in school s. Equivalently, $\hat{V}_{corrected}(\hat{\zeta}_{cgs})$ is our measure of interest when discussing the distribution of teacher effects.

[26] 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.

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

4.3.4 Sorting of Students to Teachers

Endogenous sorting of students to teachers can introduce bias to value-added estimates (Chetty, Friedman, and Rockoff 2014; Rothstein 2010; Goldhaber and Chaplin 2015; Kinsler 2012). 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 random-assignment years to years with business as usual assignment for the same set of teachers, to get a sense of the severity of the bias due to sorting.

4.4 Impact of NULP on the Distribution of Value-Added

To estimate the effects of the NULP on the distribution of value-added estimates, we move away from simply reporting estimates in the control schools but instead calculate $\hat{V}_{corrected}(\hat{\gamma}_{cgs,t})$ and $\hat{V}_{corrected}(\hat{\zeta}_{cgs})$ for the reduced-cost and the full-cost schools as well. Note that because the NULP rolled out across years as described above, some years and grades in treatment schools, teachers did not directly receive training and support - although they may have had training in the past or would have had NULP materials in their classrooms. We pool across NULP intervention arms for our estimates and provide sensitivity analyses for teachers who directly received the treatment in a given year.

4.5 Correlation with Teacher Characteristics

To understand which teachers are associated with higher value-added, we estimate the following equation:

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

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

5 Results

5.1 Classroom and Teacher Value-Added in Uganda

We first begin by presenting our estimates of classroom and teacher value-added in control group schools. Table 2 presents our estimates of teacher and classroom effects using students in the two 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 assessments. We present our estimates with and without corrections for sampling variance and present school-level cluster-bootstrapped confidence intervals in square brackets.

[Table 2 about here.]

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 0.35 SDs; for teacher effects, the estimate is 0.29 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 λ_{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 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 0.33(0.23) 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 e.g. 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 0.53(0.45) SDs (Panel B, Columns 1 and 2). After purging the school effects, the estimates are between 43 and 58 percent smaller; our preferred

estimated variance of classroom(teacher) value-added are 0.30(0.19) SDs (Panel B, Columns 3 and 4).

Local language teacher value-added is highly correlated with English: the estimates for the two subjects (after purging school effects) have a correlation coefficient of 0.73. 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).

5.2 Random Assignment of Students to Classrooms

To investigate the degree of bias due to sorting of students to classes, we re-estimate classroom effects using a different sample of teachers—those who teach in either random assignment years, 2013, 2016, and 2017 or business-as-usual years (2014 and 2015). Column 1 of Table 3 presents the results from random assignment years and Column 2 presents the results using only business-as-usual assignment years. This enables us to compare estimates using data with business-as-usual assignment to those that use only the random assignment years for the same set of teachers. For comparison, Column 3 presents the results for same subset of teachers but using all years of data available. We present results for classroom effects only because we have too few (five) teachers teaching in two random assignment years as well as two business as usual assignment years. The same limitation applies for English as we lose all grade-1 teachers and thus we only present results for local language.

[Table 3 about here.]

For Leblango, the difference in variance of classroom effects across the three cuts of data is negligible. The fact that our estimates do not vary greatly across assignment regimes is in line with balance tests in student characteristics. Our tests for sorting in the business-as-usual years parallel the tests for sorting in the random-assignment years discussed above and presented in Appendix Figure A1 (Horvath 2015). Only 5 percent of the schools had classrooms with statistically significant baseline differences between streams. Utilizing the fact that we follow the same teachers across both random assignment years and business-as-usual years allows us to test if the classroom effects obtained under business as usual assignment can predict classroom effects obtained under random assignment. The result of this test shows a positive and statistically significant correlation of 0.35.[28]

On the other hand, head teacher surveys conducted in 2017 asked about pupil assignment and find 18 percent of head teachers report sorting on student ability. This is somewhat

[28] Note that this test can only be done for a subset of the teachers (90 teachers teaching in both random assignment and business-as-usual years.)

different than what we find in our data. Of the remaining schools, 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.[29] It is not clear based on our evidence exactly how a head teacher would assess student ability and if this is done based on actual student assessments.

5.3 Impact of the NULP on the Distribution of Value-Added

In Table 4, 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 in the control group, and so simply replicate the results in Columns 3 and 4 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.

[Table 4 about here.]

For Leblango, the program increases the variance of classroom and teacher effects. The corrected standard deviation of classroom effects increases by 20 and 30 percent in Leblango reduced- and full-cost program schools, respectively. The estimated increases in the standard deviation of teacher effects due to the program are similar in percentage terms: 39 and 52 percent for local language, reduced- and full-cost, respectively. The estimates barely change for English. To formally test the difference between the study arms we bootstrap the difference between arms and examine the fraction of resamples for which the difference is zero or smaller.[30] Based on this test we can reject the null hypothesis that the local-language reading classroom and teacher effects have equal variances in the control group and the full-cost program schools.

5.4 Who are the Most Effective Teachers?

The finding that a highly effective teacher-training program increases the spread of teacher effectiveness in Leblango means that some teachers improve more than others. For policy purposes it is interesting to figure out if the most effective teachers have any observed

[29] Several head teachers also reported sorting students randomly, based on willingness to learn, height, disability, by gender of the teacher or student age, and alphabetically.

[30] Formally, we calculate the difference of in SD of teacher and classroom effects between the control and full-cost schools; this is done for each bootstrap sample (thus 1000 differences). Then we compute the 2.5^{th} and 97.5^{th} percentile of the distribution of this difference which we use as the confidence interval of the difference. The bootstrapped differences of the SD of the classroom and teacher effects are strictly positive and the 95% confidence intervals are [0.08;0.21] and [0.06;0.17], respectively

characteristics in common. Using data from the teacher surveys, we describe how teacher characteristics correlate with value-added estimates in Table 5 and allow this to vary by treatment arm. 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 and looking at the R-squared our covariates can explain between one to six percent of the variation in value-added.

[Table 5 about here.]

We can also compare the coefficients across treatment arms. The starkest difference appear for the indicator variable of having less than five years of experience - which increases notably from the control group to the reduced-cost version, and again increases among the full-cost version teachers. These results suggest an important interaction between years of experience and the NULP training - perhaps that those with fewer years of experience may be most amenable to new ways of teaching.

Another way to gain knowledge on which teachers benefit from the program is to investigate if the NULP substantially changed rankings of the teachers. 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 relatively 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 have as her counterfactual the median teacher in the control-group distribution. To test an implication of the rank preservation assumption we follow (Bitler, Gelbach, and Hoynes 2005; 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 represents 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 surely incorrect assumption of independence of the different tests, we would expect about two or

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

three rejections. For Leblango, we obtain zero rejections when using the classroom effect estimates or teacher effect estimates. We thus we cannot reject the null of zero differences in quartile means between the control and full-cost. Our evidence is therefore consistent with the theory that the treatment had rank-preserving effects on teacher value-added.

[Table 6 about here.]

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.

5.5 Sensitivity Analyses

5.5.1 Control Group Value-Added Estimates

We present several 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 specific 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, the teachers in our Classroom Effects Sample and the Teacher Effects Sample differ somewhat from each other (56% of teachers in the Classroom Effects Sample are not in the Teacher Effects Sample). To test the potential effect of this difference in sample, Appendix Table A10 presents the equivalent estimates of classroom and teacher effects conditioning on a teacher being in the Teacher Effects Sample. The results are similar to those in Table 2.

Next, our preferred estimates in Table 2 condition on observations with least five students per teacher. 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 teachers with a small number of students on our results by re-estimating our results from Table 2, omitting teachers with fewer than 10 or 15 students. Appendix Table A11 shows that excluding classrooms with fewer than 10 or 15 sampled students per teacher barely changes the estimated variance of classroom effects.

We next address the fact that we impute missing student covariates – age, gender or prior test score – to avoid losing student-year observations. Appendix Table A12, Columns 1 and 2 presents the estimates without imputing the covariates—in other words, we omit any student-year observations with missing covariates. The variances of the classroom and teacher effects differ only slightly from 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 test scores available. Thus, the estimates in Table 2 involve imputing grade-one baseline test scores to zero, which (for consistency) we do for all first-grade students. Columns 3 and 4 of Appendix Table A12 present results where we instead omit all first-grade students from the estimates. The variances of the classroom effects are only slightly affected relative to those in Table 2, but for the teacher effect for local language drops somewhat. The variance of the teacher effect for local language now resembles the variance for English suggesting that the difference seen in Table 2 may be due to missing grade one students in English rather than fundamentally different effects across subjects.

As a final sensitivity test, we present the results when purging year-specific school effects as opposed to overall school effects in Appendix Table A13. This matters when we purge the school effects from the classroom effects as the SD of the classroom effects drops compared to Table 2. The most likely reason for this difference is that the year-specific school effect is estimated with more measurement error (especially in the early years) compared to the overall school effect.[32] For the teacher effects this restriction doesn't matter as these are calculated across years and thus we subtract the overall school effect by construction.

5.5.2 Effects of the NULP

As described above, the NULP intervention was only implemented for certain grade levels in certain years; see Appendix Table A1. To address sensitivity to this feature we perform three sensitivity tests in Appendix Tables A14, A15 and A16.

First, we omit data collected in 2017 as the NULP was only implemented from 2013 to 2016 (Appendix Table A14). This leaves the classroom effect estimates nearly unchanged, but reduces the estimated variance for the teacher effects; this difference may arise 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 (Appendix Table A15). These estimates show similar patterns to Table 4.

[32] Recall that the number of grade levels included in the study increases over time (see Appendix Table A3)

Finally, we show sensitivity for omitting grade one students as we did see some sensitivity to that in Appendix Table A12. Appendix Table A16 presents the results and shows a similar pattern of increased variance for local language but not for English albeit the levels are lower for local language.

6 Conclusion

Using five years of data from students and teachers combined with a randomized evaluation of a literacy program we find substantial variation in teacher effectiveness. This variation increases when teachers are exposed to teacher training and support that increased student learning. Our findings have at least two implications.

First, despite the fact that learning levels are generally low in Uganda we find that some teachers are more effective in increasing learning than others. This points to a potential for learning from the most effective teachers in this setting. Unfortunately, we do not make a lot of headway in understanding who the most effective teachers are or what they do. One interesting avenue could be to collect detailed data on teacher attendance to see if lack of attendance can explain the variation in teacher effectiveness, or more detailed data on teacher behavior within the classroom as in Araujo et al. (2016).

Second, the NULP resulted in massive average gains in student learning. We find that the NULP also increases the variance of teacher effectiveness – by making the most effective teachers even more effective. This implies that educational interventions might increase inequality in education as more skilled teachers are better able to make use of their training. This result suggests that an important avenue for future research is to look at how to better reach less-effective teachers.

The variance in teacher value-added is usually interpreted as the scope for improving learning outcomes through teachers. Yet, comparing across very different settings may not make a lot of sense - in low income countries even the best teacher might have scope to improve, and a low variance of value-added does not necessarily suggest that there is no potential for teachers to help student learning. Even in settings with low teacher value-added, interventions that support and train teachers have the ability to improve teachers at all levels of quality. We show that it is possible to impact (and even increase) the variance of teacher value-added through educational interventions. This raises important questions about how to best help support low-quality teachers and calls for additional research on equity in teacher interventions.

We show that - despite the modest reports from head teachers that they sort students by ability, sorting is not an issue for estimation in this setting. More descriptive work could

shed light on classroom dynamics in Africa, beyond our understanding on tracking (Duflo, Dupas, and Kremer 2011) and "teach at the right level" (TARL) (Banerjee et al. 2016).

Finally, 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 will do little to improve educational outcomes. 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.

Our approach - to combine estimates of classroom and teacher value-added with a randomized teacher-focused intervention, allows us to understand the causal effects of teacher training and support. Rather than offering idle conjecture regarding the effect of moving teachers up the distribution of quality, we can observe the distribution shift.

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

References

Aaronson, Daniel, Lisa Barrow, and William Sander (2007). "Teachers and student achievement in the Chicago public high schools". *Journal of labor Economics* 25.1. Publisher: The University of Chicago Press, pp. 95–135.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady (2016). "Teacher Quality and Learning Outcomes in Kindergarten". *The Quarterly Journal of Economics* 131.3, pp. 1415–1453. ISSN: 0033-5533. DOI: 10.1093/qje/qjw016.

Azam, Mehtabul and Geeta Gandhi Kingdon (2015). "Assessing teacher quality in India". *Journal of Development Economics* 117, pp. 74–83. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2015.07.001.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton (2016). *Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India*. Working Paper 22746. Series: Working Paper Series. National Bureau of Economic Research. DOI: 10.3386/w22746.

Bau, Natalie and Jishnu Das (2020). "Teacher value added in a low-income country". *American Economic Journal: Economic Policy* 12.1, pp. 62–96.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes (2005). *Distributional Impacts of the Self-Sufficiency Project*. Working Paper 11626. National Bureau of Economic Research. DOI: 10.3386/w11626.

Blackmon, William K (2017). *Using a value-added model to measure private school performance in Tanzania*. Georgetown University.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane (2017). "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa". *Journal of Economic Perspectives* 31.4, pp. 185–204. ISSN: 0895-3309. DOI: 10.1257/jep.31.4.185.

Buddin, Richard and Gema Zamarro (2009). "Teacher qualifications and student achievement in urban elementary schools". *Journal of Urban Economics* 66.2. Publisher: Elsevier, pp. 103–115.

Buhl-Wiggers, Julie, Jason Kerwin, Juan Sebastián Muñoz, Jeffrey Smith, and Rebecca Thornton (2022). "Some Children Left Behind: Variation in the Effects of an Educational Intervention". *Journal of Econometrics* Forthcomming.

Buhl-Wiggers, Julie, Jason 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* 126.4, pp. 1593–1660. ISSN: 0033-5533, 1531-4650. 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* 104.9, pp. 2593–2632. ISSN: 0002-8282. DOI: 10.1257/aer.104.9.2593.

Crawfurd, Lee and Phil Elks (2019). "Testing the feasibility of a value-added model of school quality in a low-income country". *Development Policy Review* 37.4, pp. 470–485. ISSN: 1467-7679. DOI: 10.1111/dpr.12371.

Deininger, Klaus (2003). "Does cost of schooling affect enrollment by the poor? Universal primary education in Uganda". *Economics of Education Review* 22.3, pp. 291–305. ISSN: 0272-7757. DOI: 10.1016/S0272-7757(02)00053-5.

Djebbari, Habiba and Jeffrey Smith (2008). "Heterogeneous Impacts in Progresa". *Journal of Econometrics* 145.1, pp. 64–80. DOI: 10.1016/j.jeconom.2008.05.012.

Dubeck, Margaret M. and Amber Gove (2015). "The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations". *International Journal of Educational Development* 40, pp. 315–322. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2014.11.004.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". *American Economic Review* 101.5, pp. 1739–1774. ISSN: 0002-8282. DOI: 10.1257/aer.101.5.1739.

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* 31.2, pp. 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*. Working Paper 20284. National Bureau of Economic Research. DOI: 10.3386/w20284.

Glewwe, Paul and Karthik Muralidharan (2016). "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications". *Handbook of the Economics of Education*. Ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 5. Elsevier, pp. 653–743. DOI: 10.1016/B978-0-444-63459-7.00010-5.

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* 8.1. Publisher: Taylor & Francis, pp. 8–34.

Goldhaber, Dan, James Cowan, and Joe Walch (2013). "Is a good elementary teacher always good? Assessing teacher performance estimates across subjects". *Economics of Education Review* 36. Publisher: Elsevier, pp. 216–228.

Gove, Amber and Anna Wetterberg (2011). *The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy*. RTI International. ISBN: 978-1-934831-08-3.

Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge (2015). "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures". *Journal of Educational and Behavioral Statistics* 40.2, pp. 190–222. ISSN: 1076-9986, 1935-1054. DOI: 10.3102/1076998615574771.

Hanushek, Eric A and Steven G Rivkin (2012). "The distribution of teacher quality and implications for policy". *Annu. Rev. Econ.* 4.1. Publisher: Annual Reviews, pp. 131–157.

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* 41.5, pp. 669–683. ISSN: 0305-7925. DOI: 10.1080/03057925.2011.581014.

Horvath, Hedvig (2015). *Classroom Assignment Policies and Implications for Teacher Value-Added Estimation*.

Jackson, C Kirabo (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina*. National Bureau of Economic Research.

Kane, Thomas J and Douglas O Staiger (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. Working Paper 14607. National Bureau of Economic Research. DOI: 10.3386/w14607.

Kim, Thomas and Saul Axelrod (2005). "Direct instruction: An educators' guide and a plea for action." *The Behavior Analyst Today* 6.2, p. 111.

Kinsler, Josh (2012). "Assessing Rothstein's critique of teacher value-added models". *Quantitative Economics* 3.2, pp. 333–362. ISSN: 1759-7331. DOI: 10.3982/QE132.

Koedel, Cory (2009). "An empirical analysis of teacher spillover effects in secondary school". *Economics of Education Review* 28.6. Publisher: Elsevier, pp. 682–692.

Kremer, Michael, Conner Brannen, and Rachel Glennerster (2013). "The Challenge of Education and Learning in the Developing World". *Science* 340.6130, pp. 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* 85.3, pp. 353–394. DOI: 10.3102/0034654314553127.

Muñoz-Chereau, B and Sally M Thomas (2016). "Educational effectiveness in Chilean secondary education: Comparing different 'value added'approaches to evaluate schools". *Assessment in Education: Principles, Policy & Practice* 23.1. Publisher: Taylor & Francis, pp. 26–52.

Oketch, Moses, Caine Rolleston, and Jack Rossiter (2021). "Diagnosing the learning crisis: What can value-added analysis contribute?" *International Journal of Educational Development* 87, p. 102507. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2021.102507.

Piper, Benjamin (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* 73.2, pp. 417–458. ISSN: 0012-9682.

Rothstein, Jesse (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement". *The Quarterly Journal of Economics* 125.1, pp. 175–214.

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

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio, and Li Feng (2012). "Value added of teachers in high-poverty schools and lower poverty schools". *Journal of Urban Economics* 72.2, pp. 104–122. ISSN: 0094-1190. DOI: 10.1016/j.jue.2012.04.004.

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* 74.5. Publisher: Wiley Online Library, pp. 629–645.

Spreen, Carol Anne and Jillian J Knapczyk (2017). "Measuring Quality beyond Test Scores: The Impact of Regional Context on Curriculum Implementation (in Northern Uganda)." *FIRE: Forum for International Research in Education*. Vol. 4. Issue: 1. ERIC, pp. 1–31.

Todd, Petra E. and Kenneth I. Wolpin (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement". *The Economic Journal* 113.485, F3–F33. ISSN: 1468-0297. DOI: 10.1111/1468-0297.00097.

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

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

World Bank (2020). "School enrollment, primary (% net)". *World Development Indicators*.

Tables

Table 1
Samples Across Study Arms

	All	Control	Reduced Cost	Full Cost
Panel A: NULP Evaluation Sample				
#Schools	128	42	44	42
#Teachers	1,382	470	485	427
#Classrooms	2,200	728	762	710
#Sampled student-year obs	75,357	24,095	26,145	25,117
#Sampled students	29,787	9,572	10,454	9,761
#Students with at least one endline test	27,943	8,948	9,799	9,196
#Student-year obs with at least one EL test	2	2	2	2
#Teachers per grade	2.24	2.30	2.20	2.24
Panel B: Students with Consecutive Tests				
#Student-year obs with endline local language	58,775	18,636	20,421	19,718
#Student-year obs with prior & endline local language	49,044	15,419	17,041	16,584
#Student-year obs with endline English	37,077	11,715	12,821	12,541
#Student-year obs with prior & endline English	27,290	8,486	9,412	9,392
Panel C: Matching Students to Teachers				
#Student-year obs matched to a teacher	58,223	18,476	20,157	19,590
#Student-year obs with 5 students per teacher	56,035	17,610	19,391	19,034
Panel D: Classroom Effects 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	27,609	8,820	9,670	9,119
#Student-year obs	56,035	17,610	19,391	19,034
Panel E: Teacher Effects 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	24,218	7,468	8,678	8,072
#Student-year obs	38,082	11,429	13,280	13,373

Notes: The 128 schools were sampled in two phases: 38 in 2013 and additional 90 in 2014. Prior test is defined as an endline test in the year before. The Classroom Effects Sample includes all students and teachers available in schools where there are at least two teachers across all years. The Teacher Effects Sample includes all students with teachers teaching at least two different years. Both the Classroom Effects Sample and the Teacher Effects Sample are based on students with two local language tests, the numbers for English are smaller as there is no test in grade 1.

Table 2
Classroom and Teacher Value-Added: Control Schools

	Including School Effects		**School Effects Purged**	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Leblango Reading	(1)	(2)	(3)	(4)
SD of effects	0.40	0.31	0.38	0.26
	[0.32,0.49]	[0.22,0.40]	[0.29,0.46]	[0.18,0.34]
Corrected SD of effects	0.35	0.29	0.33	0.23
	[0.26,0.45]	[0.20,0.38]	[0.23,0.43]	[0.15,0.32]
Observations (student-years)	17,610	11,429	17,610	11,429
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	99	43	99
Panel B: English Reading				
SD of effects	0.55	0.46	0.32	0.22
	[0.40,0.70]	[0.35,0.58]	[0.30,0.35]	[0.20,0.24]
Corrected SD of effects	0.53	0.45	0.30	0.19
	[0.37,0.68]	[0.33,0.57]	[0.27,0.32]	[0.16,0.22]
Observations (student-years)	10,880	6,115	10,880	6,115
Students	5,675	4,359	5,675	4,359
Teachers	284	99	284	99
Classrooms	390	211	390	211
Schools	42	40	42	40
Pupils per classroom/teacher	37	74	37	74

Notes: The Classroom Effects Sample includes teachers available in schools where there are at least two teachers across all year while the Teacher Effects 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. Columns 1 and 2 show the classroom and teacher effects when including school effects and Columns 3 and 4 show these when school effects are purged by subtracting off the school mean. Control schools (N=42) did not receive the NULP intervention.

Table 3

Comparison Between Random Assignment and Business-as-Usual Assignment Using the Same Sample of Teachers

	Classroom Effects		
Leblango Reading	Random Assignment Years (1)	Business-as-Usual Assignment Years (2)	All Years (3)
Corrected SD of effects	0.24 [0.20,0.28]	0.20 [0.16,0.32]	0.24 [0.20,0.29]
Observations (student-years)	3611	3748	7359
Students	3400	2985	5540
Teachers	90	90	90
Classrooms	111	117	228
Schools	30	30	30
Pupils per classroom/teacher	41	48	44

Notes: Column 1 includes only random assignment years (2013, 2016 and 2017), Column 2 includes only business as usual assignment years (2014 and 2015) and Column 3 includes all years (2013-2017). All results conditioning on teachers teaching in both random assignment years as well as business as usual years. 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. 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
Heterogeneity of Value-Added by NULP Study Arm

	Classroom Effects			Teacher Effects		
Panel A: Leblango Reading	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD of effects	0.33	0.41	0.47	0.23	0.32	0.35
	[0.25,0.41]	[0.33,0.48]	[0.41,0.52]	[0.16,0.31]	[0.24,0.40]	[0.28,0.41]
Observations (student-years)	17,610	19,391	19,034	11,429	13,280	13,373
Students	8,820	9,670	9,119	7,468	8,678	8,072
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	43	42	41	78	80	83
Panel B: English Reading						
Corrected SD of effects	0.30	0.32	0.31	0.19	0.24	0.19
	[0.25,0.34]	[0.26,0.38]	[0.27,0.36]	[0.14,0.25]	[0.17,0.30]	[0.15,0.24]
Observations (student-years)	10,880	11,945	11,950	6,115	6,977	7,218
Students	5,675	6,131	5,975	4,359	4,912	4,998
Teachers	284	297	278	99	100	111
Classrooms	390	416	390	211	233	229
Schools	42	44	42	40	44	41
Pupils per classroom/teacher	37	38	39	55	59	61

Notes: 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. To test the difference between the control and full-cost results we compute the difference of the SDs for each bootstrap sample; this gives us the 95% confidence intervals of the differences. These confidence intervals are strictly positive for both the classroom ([0.08;0.21]) and teacher ([0.06;0.17]) effects.

Table 5
Teacher Value-Add Correlation with Teacher Characteristics

	Classroom Effects			Teacher Effects		
Panel A: Leblango EGRA	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
	(1)	(2)	(3)	(4)	(5)	(6)
$\geq$ Bachelor (1=Yes)	-0.090^{**}	-0.023	-0.056	-0.074	-0.025	-0.116
	(0.044)	(0.055)	(0.082)	(0.048)	(0.068)	(0.093)
Female (1=Yes)	-0.069^{*}	-0.089^{*}	-0.038	-0.001	-0.052	-0.002
	(0.041)	(0.046)	(0.054)	(0.051)	(0.045)	(0.061)
$<$ 5 yrs of experience (1=Yes)	-0.129	0.145	0.225	0.084	0.228	0.367
	(0.184)	(0.138)	(0.268)	(0.311)	(0.190)	(0.264)
yrs of experience	-0.002	-0.007^{**}	-0.001	0.001	-0.001	-0.004
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
$<$ 5 yrs of experience (1=Yes)	0.044	-0.006	-0.048	-0.026	-0.050	-0.103
yrs of experience	(0.060)	(0.058)	(0.081)	(0.094)	(0.109)	(0.073)
Observations	470	524	501	132	154	149
R-squared	0.021	0.039	0.009	0.013	0.023	0.052
Panel B: English EGRA						
$\geq$ Bachelor (1=Yes)	-0.076^{*}	0.009	-0.041	-0.000	0.039	0.061
	(0.045)	(0.052)	(0.060)	(0.069)	(0.059)	(0.063)
Female (1=Yes)	-0.078^{**}	-0.070^{*}	-0.058	0.003	-0.015	-0.025
	(0.038)	(0.039)	(0.041)	(0.046)	(0.049)	(0.044)
$<$ 5 yrs of experience (1=Yes)	-0.006	0.019	0.191^{*}	0.269	0.062	0.135
	(0.130)	(0.124)	(0.111)	(0.256)	(0.160)	(0.138)
yrs of experience	0.005	-0.003	-0.006^{*}	0.007	-0.002	-0.011^{***}
	(0.003)	(0.003)	(0.003)	(0.005)	(0.006)	(0.003)
$<$ 5 yrs of experience (1=Yes)	0.012	0.040	-0.037	-0.051	0.036	-0.034
yrs of experience	(0.044)	(0.044)	(0.029)	(0.084)	(0.068)	(0.034)
Observations	310	338	321	87	89	98
R-squared	0.034	0.031	0.050	0.050	0.049	0.181

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 6
Tests of Rank Preservation

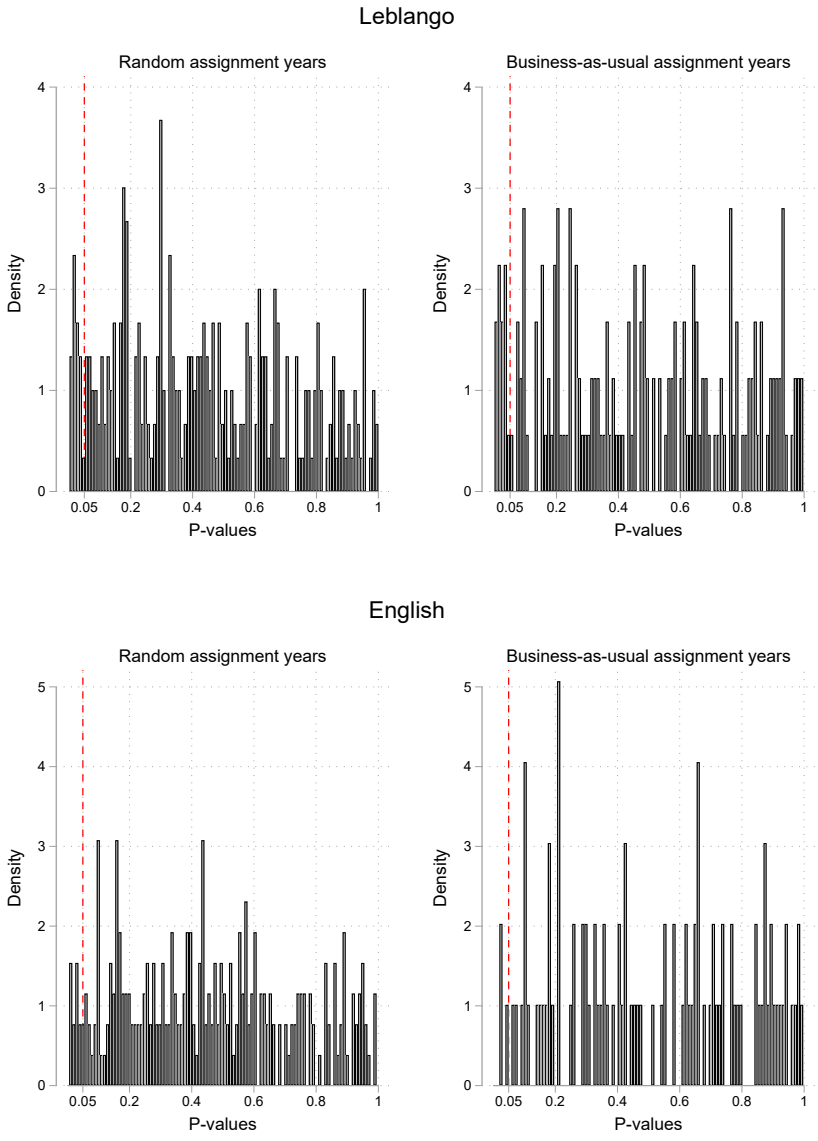
Leblango Reading	**Classroom effects**			
	Age	Gender	Experience	Schooling
	(1)	(2)	(3)	(4)
First quartile of CVA	-2.289	0.026	-1.280	0.003
	[-2.481,2.582]	[-0.137,0.147]	[-2.243,2.383]	[-0.107,0.113]
Second quartile of CVA	2.149	-0.126	1.830	-0.043
	[-2.336,2.452]	[-0.122,0.128]	[-2.255,2.488]	[-0.105,0.113]
Third quartile of CVA	-0.186	0.017	-1.183	0.031
	[-2.284,2.238]	[-0.136,0.133]	[-2.243,2.091]	[-0.105,0.108]
Fourth quartile of CVA	-0.162	0.056	-0.202	-0.008
	[-2.459,2.277]	[-0.143,0.141]	[-2.160,2.078]	[-0.107,0.115]
Observations	569	600	563	600
	Teacher Effects			
First quartile of TVA	-1.917	0.185	-0.954	0.053
	[-3.334,3.441]	[-0.201,0.207]	[-3.043,3.247]	[-0.126,0.120]
Second quartile of TVA	2.868	-0.088	0.283	-0.035
	[-2.954,2.791]	[-0.237,0.211]	[-3.071,2.853]	[-0.195,0.166]
Third quartile of TVA	-1.158	-0.016	0.899	-0.022
	[-3.897,3.710]	[-0.212,0.196]	[-3.347,3.281]	[-0.155,0.142]
Fourth quartile of TVA	-0.101	-0.103	-2.409	0.054
	[-2.942,3.024]	[-0.170,0.165]	[-2.661,2.876]	[-0.131,0.138]
Observations	284	291	281	291

Notes: Dependent Variable: Difference between Full-Cost and Control in teacher characteristics. Bootstrapped 95%-confidence intervals are in squared brackets. All regressions control for stratification cell fixed-effects. $***p < 0.01$, $**p < 0.05$, $*p < 0.1$. CVA=Classroom Value Added (using the Classrooom Effects Sample and TVA=Teacher Value Added (using the Teacher Effects Sample).

A Online Appendix

A.1 Online Appendix Figures

Appendix Figure A1
Horvarth (2015) Test



Notes: : This figure presents a distribution of p-values from regressing baseline test scores on classroom dummies within each year, school and grade-level (Horvath 2015).

A.2 Online Appendix Tables

Appendix Table A1

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 receiving NULP	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					
Grades assessed	Grade 1	Grades 1-2	Grades 1-3	Grades 1-4	Grades 3-5
Leblango reading tests (all grades)	Baseline & Endline	Baseline & Endline	Endline	Endline	Endline
English oral tests (grade-one only)	Baseline & Endline	Baseline & Endline	Endline	Endline	
English reading tests (grades > 1)		Baseline & Endline	Endline	Endline	Endline

Notes: * In 2017, Grade 4 teachers in the treatment arms received a different version of the NULP that involved being mentored by "mentor teachers".

Appendix Table A2

Number of Students per School Sampled by School Sample and Year

	2013	2014	2015	2016
Panel A: Original 38 schools sampled in 2013				
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 (Baseline sample)			30 grade-1 students	
Cohort 3 (Endline sample)				30 grade-2 students
Cohort 4				60 grade-1 students
Panel B: New 90 schools sampled in 2014				
Cohort 2 (Baseline sample)		80 grade-1 students		
Cohort 2 (Endline sample)		20 grade-1 students		
Cohort 3 (Baseline sample)			30 grade-1 students	
Cohort 3 (Endline sample)				30 grade-2 students
Cohort 4				60 grade-1 students

Notes: This table describes the sampling strategy of students for each year and grade.

Appendix Table A3

Tests Used to Estimate Value-Added

		2013	2014	2015	2016	2017
Panel A: Leblango Reading						
Grade 1	Prior Score:	0	0	0	0	
	Current Score:	Endline 2013	Endline 2014	Endline 2015	Endline 2016	
Grade 2	Prior Score:		Endline 2013	Endline 2014	Endline 2015	
	Current Score:		Endline 2014	Endline 2015	Endline 2016	
Grades 3-5	Prior Score:			Endline 2014	Endline 2015	Endline 2016
	Current Score:			Endline 2015	Endline 2016	Endline 2017
Panel B: English Reading						
Grade 1			*Not assessed in English reading*			
Grade 2	Prior Score:		Endline 2013 (oral)	Endline 2014 (oral)	Endline 2015 (oral)	
	Current Score:		Endline 2014	Endline 2015	Endline 2016	
Grades 3-5	Prior Score:			Endline 2014	Endline 2015	Endline 2016
	Current Score:			Endline 2015	Endline 2016	Endline 2017

Notes: This table presents which assessments are used to estimate value-added for each year, grade, and subject.

Appendix Table A4

Descriptive Statistics across Treatment Arms and Samples

	Classroom Effects Sample				Teacher Effects Sample			
	Control	Reduced-Cost	Full-Cost	p-value from F-test between study arms	Control	Reduced-Cost	Full-Cost	p-value from F-test between study arms
Panel A: Students	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (%)	0.496	0.507	0.496	0.77	0.491	0.515	0.497	0.41
Age	8.940	8.981	8.984	0.83	8.638	8.678	8.630	0.85
Leblango test score prior to intervention (standardized PCA index)	2.108	2.172	2.101	0.53	2.180	2.266	2.026	0.17
Panel B: Teachers								
Female (%)	0.464	0.457	0.405	0.09	0.569	0.506	0.468	0.08
Age	39.852	40.519	39.509	0.20	40.030	41.743	39.318	0.17
Yrs of experience	14.273	14.451	14.146	0.60	14.462	15.416	14.336	0.59
<5 yrs of experience	0.092	0.091	0.105	0.85	0.073	0.087	0.110	0.42
Yrs of education	14.758	14.601	14.590	0.77	14.706	14.631	14.565	0.93
Below bachelor	0.806	0.845	0.848	0.72	0.818	0.819	0.844	0.92
Bachelor or above	0.194	0.155	0.152	0.72	0.182	0.181	0.156	0.92
#Teachers with characteristics data	282	308	281		132	154	149	

Notes: The Classroom Effects Sample includes teachers available in schools where there are at least two teachers across all year while the Teacher Effects Sample includes teachers available in at least two different years between 2013 and 2017. Columns 4 and 8 present the p-value from an F-test testing the difference across treatment arms.

Appendix Table A5

Correlation between Student Attrition and Student Characteristics

Student characteristics	**Control** (1)	**Reduced-cost** (2)	**Full-cost** (3)	**All** (4)
Female (1=Yes)	0.001	0.012***	0.002	0.002
	(0.005)	(0.004)	(0.004)	(0.005)
Female × Full-cost				0.000
				(0.007)
Female × Reduced-cost				0.010
				(0.007)
Age	-0.015***	-0.013***	-0.012***	-0.013***
	(0.002)	(0.002)	(0.002)	(0.002)
Age × Full-cost				0.002
				(0.003)
Age × Reduced-cost				-0.003
				(0.003)
Full-cost program				-0.051
				(0.032)
Reduced-cost program				-0.018
				(0.028)
Observations	23,669	25,678	24,686	74,033
Adjusted R-squared	0.059	0.060	0.047	0.054

Notes: Attrittion defined within years (ie. present at baseline but missing at endline within the same year). *,**,*** denotes statistically significance at the 10, 5 and 1 percent-level, respectively.

Appendix Table A6

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	266	291	272	829
Adjusted R-squared	0.044	0.034	0.013	0.030

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

Appendix Table A7
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 × Full-cost				0.011
				(0.007)
> Bachelor × 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 × Full-cost				-0.015**
				(0.007)
< 5 yrs of experience × 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

Appendix Table A8
Correlation between Student and Teacher Characteristics

Dependent variable:	**Baseline Leblango Reading**			
	Control	**Reduced-cost**	**Full-cost**	**All**
	(1)	(2)	(3)	(4)
Female	0.021	0.089*	-0.138**	-0.032
	-0.049	-0.048	-0.06	-0.036
Age	-0.001	0.007	-0.012**	-0.001
	-0.003	-0.005	-0.005	-0.003
Bachelor or above	-0.029	0.042	0.016	0.000
	-0.029	-0.077	-0.095	-0.043
<5 yrs of experience	-0.024	-0.095	0.052	-0.002
	-0.061	-0.086	-0.127	-0.071
Yrs of experience	0.003	-0.011*	0.004	-0.001
	-0.003	-0.006	-0.006	-0.003
Observations	12852	14468	14552	41872
Adjusted R-squared	0.045	0.105	0.127	0.106

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

Appendix Table A9

Robustness Estimates of Teacher Value-Added: Using Alternative Outcomes, Control Schools

	Simple Index		Gain Score Model	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Leblango Reading	(1)	(2)	(3)	(4)
Corrected SD of effects	0.33	0.24	0.38	0.28
	[0.32,0.52]	[0.22,0.40]	[0.37,0.57]	[0.27,0.45]
Observations (pupil-by-year)	17610	11429	17610	11429
Pupils	8820	7468	8820	7468
Teachers	365	146	365	146
Classrooms	568	347	568	347
Schools	42	40	42	40
Pupils per classroom/teacher	43	99	43	99
Panel B: English Reading				
Corrected SD of effects	0.33	0.22	0.26	0.17
	[0.32,0.44]	[0.19,0.26]	[0.23,0.30]	[0.05,0.18]
Observations (pupil-by-year)	10880	6115	10880	6115
Pupils	5675	4359	5675	4359
Teachers	284	99	284	99
Classrooms	390	211	390	211
Schools	42	40	42	40
Pupils per classroom/teacher	37	74	37	74

Notes: The Classroom Effects Sample includes teachers available in schools where there are at least two teachers across all year while the Teacher Effects Sample includes teachers available in at least two different years between 2013 and 2017. Columns 1 and 2 present estimates of classroom and teacher effects when using an alternative method for constructing the test score index (the mean of the standardized components). Columns 3 and 4 present estimates of classroom and teacher effects when using a gain score model. 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.

Appendix Table A10
Classroom and Teacher Value-Added Estimates: Same Sample of Teachers, Control Schools

Panel A: Leblango Reading	Classroom Effects (1)	Teacher Effects (2)
Corrected SD of effects	0.28	0.24
	[0.22,0.35]	[0.16,0.32]
Observations (pupil-by-year)	11,429	11,429
Pupils	7,468	7,468
Teachers	146	146
Classrooms	347	347
Schools	40	40
Pupils per classroom/teacher	45	99
Panel B: English Reading		
Corrected SD of effects	0.28	0.20
	[0.25,0.31]	[0.17,0.24]
Observations (pupil-by-year)	6,115	6,115
Pupils	4,359	4,359
Teachers	99	99
Classrooms	211	211
Schools	40	40
Pupils per classroom/teacher	38	74

Notes: All estimates conditioning on teachers being in the Teacher Effects Sample. 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 purged of school effects by subtracting off the school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A11

Robustness Estimates of Teacher Value-Added: Restricting to Classes with Minimum of 10 or 15 Students, Control Schools

	Minimum of 10 Students		**Minimum of 15 Students**	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Leblango Reading	(1)	(2)	(3)	(4)
Corrected SD of effects	0.32	0.24	0.32	0.24
	[0.31,0.49]	[0.22,0.39]	[0.30,0.49]	[0.22,0.41]
Observations (pupil-by-year)	17351	10665	16851	9077
Pupils	8791	7017	8647	7017
Teachers	327	125	293	100
Classrooms	529	302	485	246
Schools	42	38	41	33
Pupils per classroom/teacher	43	102	44	106
Panel B: English Reading				
Corrected SD of effects	0.30	0.17	0.28	0.16
	[0.29,0.36]	[0.13,0.23]	[0.26,0.35]	[0.14,0.23]
Observations (pupil-by-year)	10693	5767	10325	4731
Pupils	5634	4143	5544	4143
Teachers	256	83	229	62
Classrooms	362	184	329	145
Schools	42	38	41	33
Pupils per classroom/teacher	38	77	39	80

Notes: Columns 1 and 2 present results from dropping classroom with less than 10 students. Columns 3 and 4 present results from dropping classrooms with less than 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. All estimates purged of school effects by subtracting off the school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A12

Robustness Estimates of Teacher Value-Added: Dropping Missing Observations or Grade One Students, Control Schools

	Dropping student-year observations with missing characteristics		**Omitting grade-one student-year observations**	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Leblango Reading	(1)	(2)	(3)	(4)
Corrected SD of effects	0.36	0.25	0.30	0.19
	[0.25,0.48]	[0.15,0.35]	[0.22,0.38]	[0.09,0.29]
Observations (pupil-by-year)	14237	9439	10880	6115
Pupils	7960	6525	5675	4359
Teachers	358	146	284	99
Classrooms	550	338	390	211
Schools	42	40	42	40
Pupils per classroom/teacher	40	92	37	74
Panel B: English Reading				
Corrected SD of effects	0.32	0.19	0.30	0.20
	[0.27,0.36]	[0.14,0.23]	[0.27,0.33]	[0.17,0.23]
Observations (pupil-by-year)	7507	4125	10880	6115
Pupils	4230	3049	5675	4359
Teachers	275	98	284	99
Classrooms	371	202	390	211
Schools	42	40	42	40
Pupils per classroom/teacher	29	53	37	74

Notes: Columns 1 and 2 present results from dropping observations with missing data. Columns 3 and 4 present results from dropping all grade 1 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. All estimates purged of school effects by subtracting off the school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A13

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

	Including School Effects		**School Effects Purged**	
	Classroom Effects	Teacher Effects	Classroom Effects	Teacher Effects
Panel A: Leblango Reading	(1)	(2)	(3)	(4)
SD of effects	0.40	0.31	0.26	0.26
	[0.32,0.49]	[0.22,0.41]	[0.19,0.32]	[0.17,0.36]
Corrected SD of effects	0.35	0.29	0.19	0.24
	[0.25,0.45]	[0.19,0.40]	[0.10,0.28]	[0.14,0.34]
Observations (student-years)	15,719	9,862	15,719	9,862
Students	8,454	6,737	8,454	6,737
Teachers	356	140	356	140
Classrooms	540	321	540	321
Schools	42	39	42	39
Pupils per classroom/teacher	39	87	39	87
Panel B: English Reading				
SD of effects	0.55	0.49	0.23	0.23
	[0.42,0.68]	[0.38,0.60]	[0.20,0.25]	[0.19,0.27]
Corrected SD of effects	0.53	0.48	0.18	0.21
	[0.39,0.67]	[0.37,0.58]	[0.15,0.22]	[0.17,0.25]
Observations (student-years)	10,880	5,921	10,880	5,921
Students	5,675	4,225	5,675	4,225
Teachers	281	94	281	94
Classrooms	390	206	390	206
Schools	42	39	42	39
Pupils per classroom/teacher	37	75	37	75

Notes: The Classroom Effects Sample includes teachers available in schools where there are at least two teachers across within each year while the Teacher Effects 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. Columns 1 and 2 show the classroom and teacher effects when including school effects and Columns 3 and 4 show these when school effects are purged by subtracting off the year-specific school mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A14

Robustness Heterogeneity of Value-Added by NULP Study Arm, 2017 Data Omitted

	Classroom Effects			Teacher Effects		
	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
Panel A: Leblango Reading	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD of effects	0.26	0.41	0.49	0.15	0.23	0.32
	[0.23,0.42]	[0.40,0.55]	[0.43,0.56]	[0.12,0.26]	[0.23,0.32]	[0.25,0.39]
Observations (pupil-by-year)	13,783	15,206	14,892	8,303	9,367	9,825
Pupils	8,579	9,441	8,991	6,515	7,323	7,092
Teachers	277	292	262	101	114	112
Classrooms	423	461	438	239	275	279
Schools	42	44	42	35	38	37
Pupils per classroom/teacher	45	43	43	34	34	36
Panel B: English Reading						
Corrected SD of effects	0.26	0.32	0.32	0.11	0.15	0.18
	[0.23,0.34]	[0.31,0.45]	[0.25,0.38]	[0.11,0.20]	[0.13,0.28]	[0.10,0.25]
Observations (pupil-by-year)	7,053	7,760	7,808	3,195	3,709	4,273
Pupils	5,036	5,467	5,472	2,968	3,373	3,719
Teachers	194	202	188	45	44	58
Classrooms	245	263	247	110	124	128
Schools	42	44	42	34	36	36
Pupils per classroom/teacher	39	39	40	24	24	30

Notes: All estimates are calculated using data between 2013 and 2016. 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.

Appendix Table A15

Robustness Heterogeneity of Value-Added by NULP Study Arm, only Treated Teachers

	Classroom Effects			Teacher Effects		
	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
Panel A: Leblango Reading	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD of effects	0.28	0.42	0.50	0.23	0.29	0.35
	[0.26,0.41]	[0.41,0.54]	[0.46,0.56]	[0.21,0.35]	[0.28,0.46]	[0.32,0.43]
Observations (pupil-by-year)	13124	14752	14753	10309	12315	12670
Pupils	7654	8753	8429	6855	8340	7902
Teachers	214	225	207	125	147	146
Classrooms	395	436	425	306	359	362
Schools	42	44	42	40	44	41
Pupils per classroom/teacher	46	44	43	33	34	35
Panel B: English Reading						
Corrected SD of effects	0.26	0.31	0.32	0.20	0.24	0.20
	[0.24,0.33]	[0.30,0.41]	[0.25,0.36]	[0.13,0.22]	[0.22,0.37]	[0.15,0.28]
Observations (pupil-by-year)	7410	8003	8093	5607	6316	6569
Pupils	4949	5283	5218	4149	4661	4763
Teachers	157	158	149	86	84	95
Classrooms	252	263	248	189	204	200
Schools	42	44	42	40	43	41
Pupils per classroom/teacher	39	39	41	29	29	32

Notes: All estimates are calculated using only the treated cohorts; P1 (2013 and 2014), P2 (2015), and P3 (2016). 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.

Appendix Table A16

Robustness Heterogeneity of Value-Added by NULP Study Arm, Dropping Grade one Students

	Classroom Effects			Teacher Effects		
	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
Panel A: Leblango Reading	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD of effects	0.30	0.35	0.39	0.19	0.27	0.26
	[0.22,0.38]	[0.27,0.42]	[0.36,0.43]	[0.09,0.29]	[0.17,0.37]	[0.22,0.29]
Observations (pupil-by-year)	10880	11945	11950	6115	6977	7218
Pupils	5675	6131	5975	4359	4912	4998
Teachers	284	297	278	99	100	111
Classrooms	390	416	390	211	233	229
Schools	42	44	42	40	44	41
Pupils per classroom/teacher	37	38	39	55	59	61
Panel B: English Reading						
Corrected SD of effects	0.30	0.33	0.32	0.20	0.25	0.20
	[0.27,0.33]	[0.28,0.38]	[0.28,0.35]	[0.17,0.23]	[0.18,0.32]	[0.13,0.26]
Observations (pupil-by-year)	10880	11945	11950	6115	6977	7218
Pupils	5675	6131	5975	4359	4912	4998
Teachers	284	297	278	99	100	111
Classrooms	390	416	390	211	233	229
Schools	42	44	42	40	44	41
Pupils per classroom/teacher	37	38	39	55	59	61

Notes: All estimates are calculated using data using grades 2 to 5. 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.