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*

December 22, 2023

Abstract

This paper uses longitudinal data from a school-based RCT to provide the first estimates of the variation in teacher effectiveness for sub-Saharan Africa. The lower bound on the SD of teacher effects is 0.09 SDs in local-language reading, 0.11 in English reading and 0.18 in math; we find no evidence of non-random sorting of students to teachers. Providing high-impact teacher training and support causes the variation in teacher effectiveness to increase by 78% in local-language reading, likely via improvements for already-effective teachers. Observed teacher characteristics are weakly correlated with both levels of, and treatment effects on, teacher effectiveness.

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.

1 Introduction

Extensive evidence shows that the most important predictor of student learning is the quality of a student's teacher (Rivkin, Hanushek, and Kain 2005; Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014; Kremer, Brannen, and Glennerster 2013; Evans and Popova 2016; Ganimian and Murnane 2014; Glewwe and Muralidharan 2016; McEwan 2015). Hence, if teachers vary substantially in their ability to contribute to student learning, one worry is that some children will be left behind (Buhl-Wiggers et al. 2022). This has led to research—mainly in high-income settings—focused on measuring the variation in teacher effectiveness, which is often interpreted as the scope for policies to improve student learning by targeting teachers (Rivkin, Hanushek, and Kain 2005; Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014). Implicit in this literature is the assumption that at least *some* teachers are teaching at their highest capacity. A low variance of teacher effectiveness thus implies a limited scope for policies targeting teachers. In low-resource settings, however, a low variance of teacher effectiveness may simply mean that most teachers are performing poorly and there may still be room for policies focused on teacher improvement. Existing research has mainly focused on providing estimates of static measures of the distribution of teacher quality, rather than studying how policies can affect the variation in effectiveness.

This paper fills that gap by first estimating the variance of teacher effectiveness in Uganda and then measuring the effect of a teacher-focused intervention on the distribution of teacher effectiveness. We first outline a theoretical framework of the production of learning to describe the relationship between the distribution of teacher effectiveness and the level of available educational resources. This framework provides a structure for understanding how interventions that provide inputs or support to teachers can affect the distribution in teacher quality. Our framework also motivates how tests of rank preservation can provide insight into which teachers benefit from education interventions. Guided by our model of learning and teacher effectiveness, we use longitudinal data from schools, teachers, and students in Uganda to estimate the first teacher value-added estimates in sub-Saharan Africa. We then estimate the causal effects of a randomized teacher focused intervention—the Northern Uganda Literacy Project (NULP)—on the distribution of teacher value-added.

Our first set of results involve estimating lower bounds on the variation of classroom and teacher effectiveness (measured as teacher value-added). These results use data from 42 control schools that did not receive the NULP from 2013 to 2017.[1] Following previous

[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. In work that post-dates our study, Oketch, Rolleston, and Rossiter (2021) estimate classroom value added in Ethiopia, but do not isolate the

research, we distinguish between "classroom effects", which are the causal effect of being in a specific classroom in a given year, and "teacher effects", which are the stable component of classroom effects that can be attributed to a given teacher. Classroom value-added is estimated using student test scores where we observe at least two teachers per school; teacher value-added is estimated for teachers who teach across multiple years in our sample. We present two main sets of estimates: within-school estimates that do not account for sorting of teachers across grade-levels, and within-school-grade estimates that do. Our within-school estimates show that a one standard deviation increase in teacher value-added improves local-language reading test scores by 0.24 SDs, English reading by 0.22 SDs, and math by 0.30 SDs. These results account for systematic sorting of teachers to schools.

We show evidence that in our setting there is systematic sorting of teachers to grades. Because of this sorting by grade, our preferred estimates focus on variation solely within a specific grade for a given school. These estimates are 0.09 SDs for local language, 0.11 for English, and 0.18 for Math. Our study also addresses the potential bias arising from non-random sorting of students to classrooms by utilizing the fact that three of the five years of data collection randomly assigned teachers to classrooms. Comparing our preferred estimates (using all five years of data) to estimates obtained from the random assignment years shows barely any difference, suggesting limited systematic student sorting.

As we discuss in our theoretical framework, the variance in teacher effectiveness may be larger or smaller in lower-income settings.[2] In our setting of northern Uganda, one of extreme poverty in a post-conflict area of sub-Saharan Africa, our within-school estimates are about twice those in the United States while the within-grade estimates are about on par with those from other settings such as Ecuador.[3] Teacher value-added is positively correlated across subjects. This suggests that teachers are broadly effective across subjects, consistent with prior literature (Koedel, Betts, et al. 2007; Loeb, Kalogrides, and Béteille

stable contributions of specific teachers.

[2] On the one hand, a difficult teaching environment may mean that good teachers are much more important than the worst teachers. On the other hand, difficult conditions may make it impossible for even the most teachers to be effective. Data from sub-Saharan Africa paint a bleak picture of the school teaching environment Bold et al. (2017) find that "essentially no public primary schools in... [Kenya, Mozambique, Nigeria, Senegal, Tanzania, Togo, and Uganda] offer adequate quality education".

[3] In American primary schools, estimates of the teacher value-added in reading are around 0.13 SDs of test scores—the average across nine studies (Hanushek and Rivkin 2012). Chetty, Friedman, and Rockoff (2014) estimate an SD of teacher value-added of 0.10 SDs among students in US schools. Araujo et al. 2016 finds a standard deviation of 0.09 SDS among kindergarten teachers in Ecuador, and Bau and Das (2020) estimate a SD of 0.06 in Pakistan. The finding of a wider distribution of teacher value-added in lower-income settings is consistent with Sass et al. (2012) who find greater variation in teacher value-added high poverty schools than lower poverty ones among students in North Carolina and Florida. Azam and Kingdon (2015) provide estimates from India that are substantially larger than ours, at 0.37 SDs, but differ in that their results are for gains over two years (corresponding to an annual gain of roughly 0.18 SDs), and they focus on teachers in secondary, rather than primary schools.

2012; Goldhaber, Cowan, and Walch 2013; Condie, Lefgren, and Sims 2014).

The second set of results in the paper involves estimating the causal effect of an educational intervention on the distribution of teacher value-added. The NULP provided intensive training and support to teachers in grades one to three for literacy instruction, with a focus on local-language reading. We utilize the random assignment of our sample schools to three treatment arms: a control that did not receive the NULP, a full-cost version of the program in which the NULP was delivered directly to teachers, and a reduced-cost version of the program in which the NULP was implemented using a cascade model of delivery in collaboration with government tutors. Both versions of the NULP resulted in massive increases in student learning: three years of exposure to the intervention causes students in full-cost program schools to score 1.35 SDs higher and students in reduced-cost program schools to score 0.78 SDs higher in local language reading (Buhl-Wiggers et al. 2018).

We present the effects of the intervention on the distribution of value-added for local language reading (rather than English or math), since this was the focus of the NULP intervention. We estimate value-added separately in each treatment arm, finding that both versions of the intervention increase the spread of the distribution of teacher effectiveness. Specifically, the intervention causes the SD of teacher effects to increase from 0.09 SDs in control schools to 0.14 SDs in reduced-cost and 0.16 SDs in full-cost schools.

Our theoretical framework shows how testing for rank preservation in teacher quality across intervention treatment arms can help provide insight into which types of teachers drive the change in the variation of teacher value-added. If the NULP is rank preserving—i.e., if teachers maintain their rank within the distribution of quality—an increase in the distribution of teacher value-added is consistent with a "skills beget skills" theory where the best teachers improve the most. In contrast, rejecting rank preservation suggests that the least effective teachers have become the most effective as a consequence of the NULP. We formally test for rank preservation following Bitler, Gelbach, and Hoynes (2005) and Djebbari and Smith (2008), and are unable to reject the null hypothesis of rank preservation. This suggests that the NULP had the largest effects amongst the most-effective teachers.

To further explore who are the most affected teachers, we also examine how teacher characteristics correlate with our estimated measure of value-added, and the gains in value-added resulting from the NULP. Prior research has typically found that the first few years of teacher experience are important predictors of value-added, with only few other successful covariates (Azam and Kingdon 2015; Slater, Davies, and Burgess 2012; Araujo et al. 2016; Bau and Das 2020). Similar to prior studies, we can explain little of the variation in effectiveness using teacher characteristics. We find limited correlation between teacher value-added and teacher characteristics such as education level, gender and experience. There is some evi-

dence that control-group teachers with more experience are more effective, but this pattern is not evident in the treatment arms.

The paper proceeds as follows: Section 2 presents a conceptual framework to think about how the distribution of teacher value-added varies with educational inputs and the potential impact of training teachers. Section 3 describes the setting and describes the NULP intervention. In Section 4 we present descriptive results on the sorting of teachers to schools, grades and classrooms, and provide a description of our analytical samples. Section 5 presents our empirical approach and estimates of teacher effectiveness. In Section 6 we present the treatment effects of the NULP on the variation in teacher effectiveness. Section 7 concludes.

2 Conceptual Framework

In this section, we present a framework that builds on the canonical "production function" model of student achievement given different levels of educational inputs (e.g. Todd and Wolpin 2003; Rivkin, Hanushek, and Kain 2005). The framework defines our notation and provides interpretive context for our empirical analyses.

2.1 Production of Student Skills

Consider a model of academic achievement for student i, in school s, with teacher j, learning in a subject k, in year t, where we measure achievement (with error) by test scores in each subject Y^k_{isjt} (i.e., $k \in \{math, reading\}$). Achievement in year t depends on the entire sequence of choices made by parents, teachers, and schools. At home, let X_{it} denote parental inputs to student i in period t, and let $X_i(t)$ denote the vector of such inputs in all periods up to and including period t. We conceive of parental inputs as including time, cognitive stimulation, health care, books, toys and so on. Achievement also depends on the student's fixed subject-specific genetic endowment (θ^k_i).

We apportion what happens at school into inputs at three levels: the teacher, the classroom, and the school as a whole. School inputs comprise those experienced by *all* students within the school such as the head teacher and their staff, the physical environment of the school and its grounds and so on. We denote school inputs in period t relevant to the production of achievement in subject k by S^k_{st}, and the corresponding vector of such inputs up to and including period t by $S_s(t)$.

Teacher inputs include how much effort the teacher put into the teaching and the level of training that the teacher has. Thus, the degree to which classroom assignment affect student skills depends on; the materials provided (M^k_{sjt}), and how the teacher assigned utilize these

materials which in turn depends on skills (L^k_{sjt}) and effort of the teacher (E^k_{sjt}). Skill and effort can be both complements and substitutes and we define teacher effectiveness as the combination of skills and effort ($J(E^k_{sjt}, L^k_{sjt})$), where teacher effectiveness is increasing with increasing levels of skills and effort. Skill formation is a dynamic process and thus depend on the entire history of family, and school inputs. Accordingly, $X_i(t)$ and $S_s(t)$ denotes vectors of all family and school inputs experienced by the student up to time t. M^k_{sjt} denotes the history of classroom materials up to time t. In addition, achievement depends on the history of classroom materials ($M^k_{sj}(t)$) and the utilization of these ($J(E^k_{sj}(t), L^k_{sj}(t))$. In addition, let a student be endowed with Finally, ϵ^k_{isjt} allows for measurement error in test scores. Accordingly, the production function can be written as follows:

$$Y^k_{isjt} = g[X_i(t), S_s(t), M^k_{sj}(t) J(E^k_{sj}(t), L^k_{sj}(t)), \theta^k_i, \epsilon^k_{isjt}] \tag{1}$$

The fundamental problem of empirically estimating Equation (1) is that data on the entire cumulative process is rarely, if ever, available. To address this challenge one approach is to use prior student achievement (Y^k_{isjt-1}) as a proxy for unobserved input histories as well as genetic endowments (Todd and Wolpin 2003). Accordingly, Equation (1) can be re-written as follows:

$$\begin{aligned} Y^k_{isjt} =& g\big\{Y^k_{isjt-1}[X_i(t-1), S_s(t-1), M^k_{sjt}(t-1), J(E^k_{sjt}(t-1), L^k_{sj}(t-1)), \theta^k_i], \\ & X_{it}, S_{st}, M^k_{sjt}, J(E^k_{sjt}, L^k_{sjt}), \eta^k_{isjt}\big\} \end{aligned} \tag{2}$$

We focus our main attention on $J(E^k_{sjt}, L^k_{sjt})$, which represents student learning attributable to being in a specific classroom and taught by a specific teacher.[4] Teacher effectiveness can vary by subject. Some teachers may have more training, experience with or skills in a particular subject, or may have more access to relevant inputs for certain subjects. Teachers may be narrowly effective at particular subjects, for example, if they have a specialized skill in a subject or substitute effort between subjects, in which case we would observe a negative correlation in teacher effectiveness across subjects. Alternatively, teachers could be broadly effective across multiple subjects in which case we would observe a positive correlation across subjects.[5]

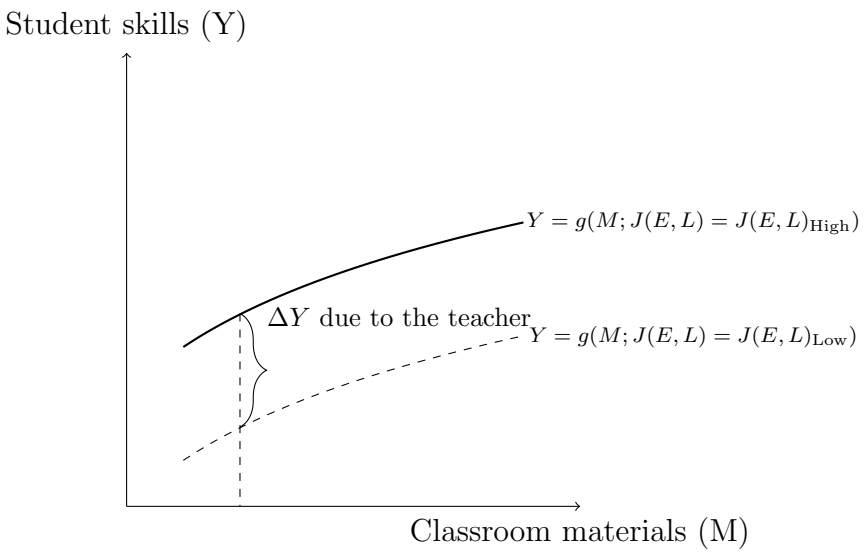
We assume that student learning is increasing in materials, teacher effort, and teacher

[4] When teachers are perfectly nested within classrooms one need multiple years of cohorts to estimate teacher effectiveness

[5] An alternative driver of a positive or negative correlation across subjects is that subjects are complements or substitutes in terms of student ability—for example, learning math may make learning physics easier, or learning how to read may make learning how to write easier. We are unable to differentiate these two channels.

skills, such that higher values of M, E and L never lead to lower student learning outcomes.[6] Figure 1 provides a simple illustration of the relationship between student skills and classroom materials at different levels of teacher effectiveness. Holding other inputs fixed student skills depend positively on classroom materials but at a decreasing rate (Brown and Saks 1981). The higher of the two curves (solid) shows what happens when teacher effectiveness is increased to a higher level; higher effectiveness is assumed to increase student achievement at every point of classroom materials. At one extreme the dashed curve could represent a case where the teacher have no skills and put in no effort leading to no additional learning. At the other extreme the solid curve could represent the case where the teacher have the highest level of skills and put in maximum effort leading to the highest amount of skills possible with a given set of materials. Thus, the difference between the dashed and solid curve can be though of as the difference in student skills when being assigned different types of teachers. The closer the curves are the more similar teachers are.

Figure 1
The Local Relationship between Student Skills and Classroom Materials by Teacher Effectiveness



$J(E^k_{sjt}, L^k_{sj})$ varies over time due to year-specific inputs. It thus captures classroom effects:

[6] This rules out a J-curve of productivity, which could happen if the use of an input is at first unproductive and then increases in productivity, which could happen with say, with new technologies, methods, or computers, for example (Kerwin and Thornton 2021). Unlike classroom inputs that exhibit diminishing returns, there is less consensus regarding the return to teacher effort. On the one hand, there may be diminishing returns, on the other hand, if skills beget skills, there may be increasing returns.

the impact on student learning of being in a specific classroom in a given year. Shocks to classroom resources (M) can include changes in physical capital such as classroom resources. When teachers teach different students each year it is possible to separate the effect from the teacher to that of the other classroom factors.

Conceptually, teacher effectiveness is value-added: the effect of having a specific teacher on a student's test scores, relative to the average teacher. We measure teacher effectiveness for a given teacher in a given year on a given subject (classroom effects), and separately estimate effectiveness after purging out year-on-year variation (teacher effects).

To estimate teacher effects (the effect of having a specific teacher on a student's test scores) we make the additional assumption that the time-varying shocks to materials, skills and peers are additively separable from the other variables that influence classroom effects, and that teacher skill is fixed over time. Formally,

$$C(M^k_{sjt}, L^k_{sj}) = T(L^k_j, M^k_j) + m^k_{jt} \tag{3}$$

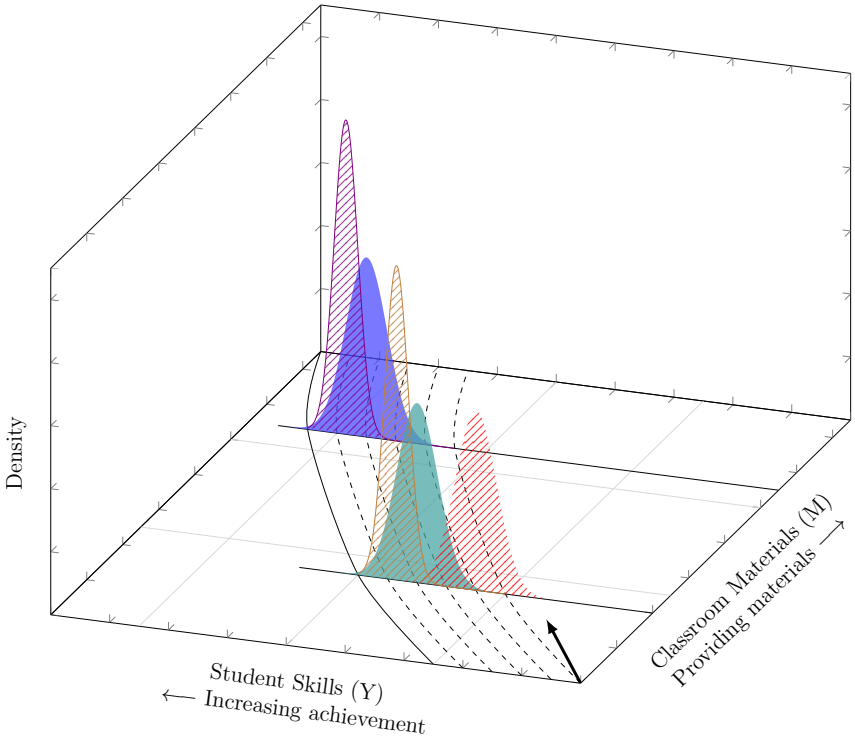
where m^k_{jt} is the mean-zero time-varying component of materials. This allows us to estimate teacher effects as the stable component of C^k_{tj}, $T^k_j = T(L^k_j, M^k_j)$. Note that in this model teacher effects can depend on time-invariant inputs such as the materials

2.2 Variation in Teacher Effectiveness

To understand the relationship between student learning, educational inputs, and teacher effectiveness, Figure 2 illustrates examples of distributions of teacher effectiveness for different levels of available educational inputs. Teacher effectiveness T for a given subject at time t is measured on the x-axis. Since learning is an increasing function of teacher effectiveness, we could also draw similar graphs with Y on the same axis. The y-axis indicates the aggregate level of inputs available to teacher j for a given subject.[7] We illustrate two kinds of inputs on the graph, materials (M) and skills (L). Low levels of inputs—represented to the left of the input axis–reflect poorer settings (such as Africa), while higher levels of inputs—represented to the right of the axis–are similar to richer settings (such as the United States). The two figures are exact replicas of one another, because the effects of both kinds of input are similar.

[7] Available inputs need not be equal to the inputs actually allocated to a student. We do not address this gap in this paper; we only know what inputs are, in theory, available.

Figure 2
A Simple Model of Teacher Effectiveness and the Production of Learning



In Figure 2, the solid line provides an upper bound for translating inputs into student learning (production possibility frontier), with increasing returns at low levels of inputs, and decreasing returns at the highest levels of inputs. Students at the lowest levels of inputs are unable to reach the highest levels of learning that are possible in resource-rich environments. Below the production possibility frontier, teachers vary in their effectiveness of translating inputs into learning. The dotted lines in Figure 2 represent different production functions for each teacher j, mapping the relationship between input levels and teacher effectiveness T. With a given set of inputs, more effective teachers have production curves that are further to the left, compared to less effective teachers with curves that are further to the right. When the productivity curves are closer together, the distribution of teacher effectiveness—represented by the normal distributions—is narrow. When productivity curves for teachers are further apart, the distribution of teacher effectiveness is wider.

It is theoretically ambiguous whether the distribution of teacher value-added will be wider or smaller in low vs. high resource settings. In an extreme case, in low resource settings, teachers may be unable to produce much learning at all. For example, it is difficult to teach students to read if there are no books or print materials. In this example, no matter

the effort, a teacher is still unable to teach students to read. This is a case in which the production curves are very tightly packed next to each other, and the variation in teacher effectiveness is close to zero. This corresponds to the left-most (brown) distribution at the bottom of the figure. The opposite case in low-resource settings is that the returns to effort and skill are even higher, because better teachers are able to find creative ways to produce student learning with limited resources. This would result in a wide distribution of teacher value-added, as in the red or green distributions at the bottom of the figure. Similarly, the variation in teacher value-added may be either narrow or wide in high-resource settings. If high levels of inputs lead to uniform success (so every teacher is extremely effective) then the distribution will be narrow; conversely, teachers may have access to many inputs but vary greatly in their ability to use technology or choose the correct inputs for their students.

In high-resource settings, the scope for improving learning comes primarily from reducing the variance between teachers (moving from the blue to the purple distribution in Figure 2), because input levels are high and the most effective teachers are already at the production possibility frontier .In low-resource settings, on the other hand, the scope for improving average learning outcomes through teachers includes both effects on the variance of the distribution of teacher effectiveness and shifts in the level of the distribution. Access to more materials (M) might raise learning outcomes for all students, but benefit the weakest teachers more than stronger ones. Similarly, better training will increase teacher skills (L), which can increase average learning outcomes but might have larger effects on weaker teachers (narrowing the spread of the distribution). It might also have larger effects for stronger teachers, which would widen the spread of the distribution.

2.3 What Happens with a Teacher-Focused Intervention?

Specifying teacher value-added as a function of materials, skills, and effort enables us to think through the effects of different policy interventions. Increased student learning can result from a variety of interventions such as increasing inputs (resulting in moving up the production curve from the bottom toward the top of the graph), or teacher training that increases skills or encourages effort (both resulting in a shift of the entire production curve to the left). Interventions such as the NULP that use a combination of inputs and training result in a diagonal movement from the bottom left corner to the top right corner (see the solid black arrow in Figure 2).

How might education interventions affect the distribution of teacher value-added? The answer depends on how teacher productivity varies with increased inputs, effort, or skill, which is ultimately related to the shape of the production curves in our model. The produc-

tion curves in Figure 2 do not cross, which implies an assumption of rank preservation as inputs increase, but this need not be the case.We first illustrate the effects of an intervention in the case of rank preservation and then turn to the case of rank inversion.

2.3.1 Rank Preservation

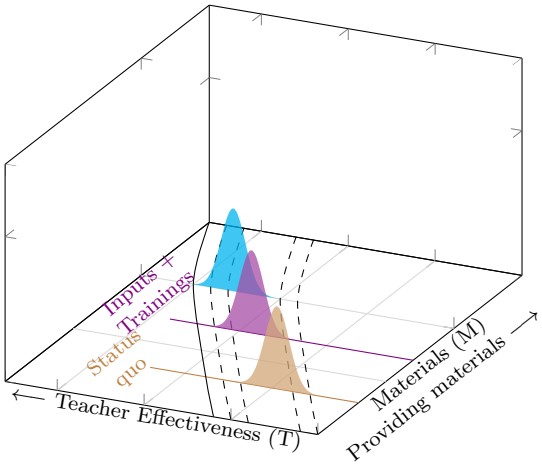
To get a sense of the effects of providing inputs and training on the distribution of teacher effectiveness, we first consider the case in which teachers maintain their rank in quality after an intervention. This assumes rank preservation, i.e. non-crossing productivity curves. There are three possible effects of providing inputs and training, on the distribution of teacher effectiveness, that we illustrate in Figure 3. These include when there is constant, decreasing, or increasing variance in the distribution of teacher effectiveness. For simplicity, our illustrations focus on the effects of changes in inputs (M). The same patterns are possible for changes in skills (L) or effort (E).

Figure 3a illustrates the case of parallel productivity curves, in which case there is constant variance of teacher effectiveness with increases in inputs. An intervention that provides more materials thus increases the effectiveness of all teachers equally. (Note that a parallel argument also applies to an intervention that increases skills or effort; the NULP increases all three). While there is a level effect in improving student learning, there is no predicted change in the distribution of teacher effectiveness. Figure 3b presents the case in which the variance of teacher effectiveness decreases in response to an intervention. If low-performing teachers have more room for improvement, teacher-focused interventions benefit low-performing teachers relatively more than already high-performing teachers i.e. the lower production curve is steeper than the upper production curve, and teachers with low skills "catch up". In this case, the distribution of teacher value-added would both shift to the right as well as narrow, as the low-performing teachers catch up with the high performing teachers. Lastly, Figure 3c presents the case of increasing variance of teacher effectiveness. This can happen if, for example, high-performing teachers are more able to take advantage of the training provided ("skills beget skills") and thus benefit relatively more than low-performing teachers.

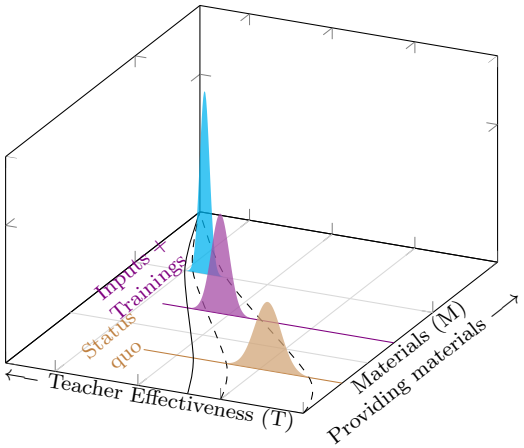
2.3.2 Rank Inversion

Teacher-focused interventions may not always be rank-preserving. One way this can happen is if an intervention specifically targets teachers at the bottom the distribution. It can also happen if the teachers who benefit the most from additional inputs are the weakest under the status quo: for example, some teachers may be very good at teaching with textbooks,

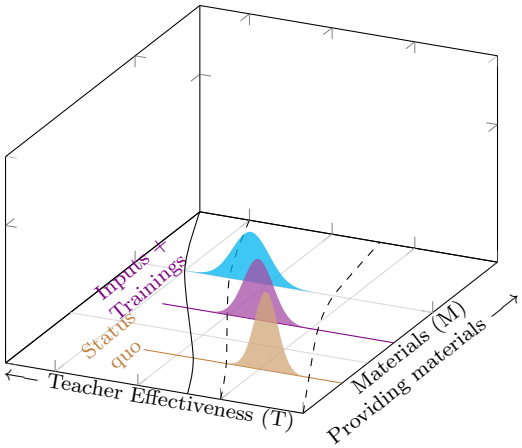
Figure 3
Impact of Increased Inputs on Learning and Teacher Effectiveness under Rank Preservation



(a) Constant Variation in Teacher Effectiveness



(b) Decreasing Variation in Teacher Effectiveness



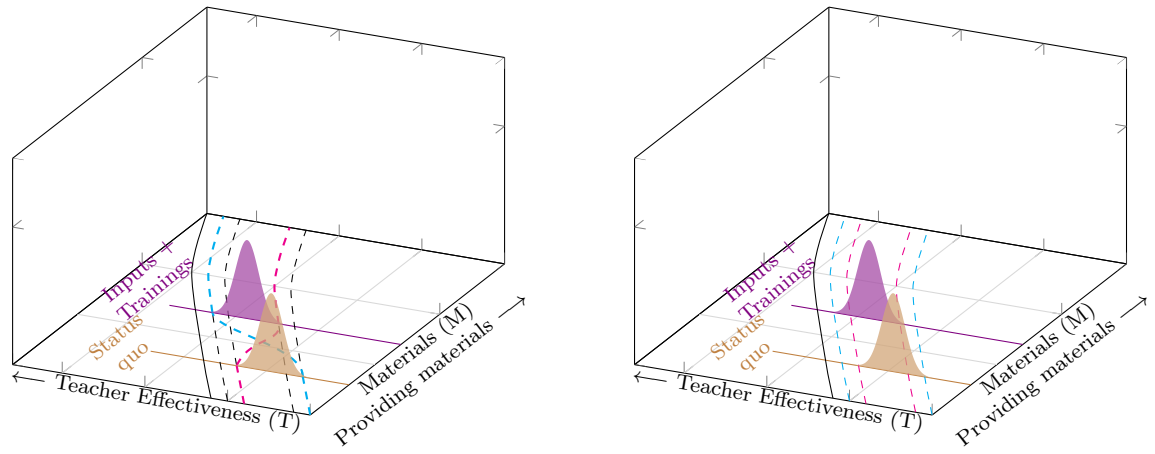
(c) Increasing Variation in Teacher Effectiveness

but perform very poorly without them. We can illustrate rank inversion in our model if either the production curves cross, such as in Figure 4a, or, if a shift in effort results in lower quality teachers "leapfrogging" higher quality teachers in Figure 4b. In both of these examples, there may be constant, narrowing, or widening productivity curves, corresponding to a constant, decreasing, or increasing variance of teacher effectiveness.

2.3.3 Predicted Effects of the NULP

The NULP program that we evaluate in this paper is a bundled intervention that provided learning inputs and teacher training and support. It and may also have affected teacher

Figure 4
Impact of Input Interventions on Learning and Teacher Effectiveness under Rank Inversion



(a) Crossing Productivity Curves **(b)** Leap-Frogging Effective Teachers

motivation and effort. Our framework provides a structure for our approach to measuring how the intervention can affect teacher value-added. In our empirical analysis, we measure the effect of the NULP on the distribution of teacher value-added, and also test for rank preservation, to shed light on which types of teachers are driving the change in the distribution of teacher value-added. The results of these tests will help inform our knowledge of the shape of the teaching production function (corresponding to one of the panels of Figure 3 or Figure 4.)

3 Setting and NULP Evaluation Design

3.1 Setting

Primary education in Uganda consists of seven years of schooling. The primary grade levels, known as Primary 1 through Primary 7, correspond to grades one through seven in the United States. While the country's net primary school enrollment rate is now above 90 percent, only about 60 percent of students transition from primary to secondary school (Deininger 2003; World Bank 2020). In addition to issues with enrollment and retention, Ugandan students face major challenges to learning. 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).[8]

[8] These statistics may even overstate student performance because schools discourage weaker students from attending in grade 7 to prepare the strongest students for the higher-stakes primary leaving exam

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). At the same time, there are many obstacles to implementing this "mother-tongue first" policy, including underdeveloped orthographies, poor instructional methodologies for reading, and a lack of relevant and adequate reading materials in many of the languages of instruction (Ssentanda, Huddlestone, and Southwood 2016; Altinyelken 2010). As a result, in practice the implementation of the policy was very limited.

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, with limited practical classroom training (Hardman et al. 2011). Teachers in Uganda receive Continuous Professional Development (CPD), in-service training intended to update classroom competencies. The CPD program is managed through primary teachers' colleges by Coordinating Center Tutors (CCTs). CCTs are typically recruited from experienced teachers and head teachers (principals). CCTs 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, they receive limited training and support, making it difficult to effectively mentor teachers (Hardman et al. 2011).

In sum, 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.

The program that we study was implemented in Northern Uganda. Of the four regions in Uganda, Northern Uganda is the poorest, with a history of marginalization. The region contains only a fifth 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).

(Gilligan et al. 2018).

The constraints that we outline for sub-Saharan or Ugandan teachers more generally, are especially challenging for teachers in Northern Ugandan schools.

3.2 NULP Intervention and Evaluation

The Northern Uganda Literacy Project (NULP) is an early-grade mother-tongue literacy program developed in response to the educational challenges facing Northern Uganda. The NULP evaluation was conducted over five years, 2013 to 2017.

3.2.1 NULP Intervention

The NULP was designed by a locally owned educational tools company, Mango Tree. It 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 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 textbooks and readers for students, and slates, chalk, and wall clocks for first-grade classrooms.

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. This approach is similar to the one used by Bridge International Academies in Kenya prior to 2020 (Gray-Lobe et al. 2022), but with a very different incentive structure for teachers. The Bridge model was run in private schools and following the scripts was mandatory. In the NULP, all teachers are employed by public schools and there are no financial or job-security reasons to follow the suggested lesson plans.

The full-cost version of the NULP consisted of the original literacy program as designed and delivered by Mango Tree and its staff. In addition, a reduced-cost NULP was implemented in some schools, following a "cascade" or "training-of-trainers" delivery model led by Ministry of Education CCTs rather than Mango Tree staff; teachers in these schools also received fewer support visits.[9]

[9] Two of the material inputs provided by the NULP—the slates and wall clocks—were provided only to a

The NULP was introduced to different grades in treated schools over the course of 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.[10] Classrooms were allowed to keep all of the Mango Tree educational materials (such as slates, primers, and readers) after they received the program, but teachers no longer received additional training or support visits.

3.2.2 Sample

There are 128 schools involved in the evaluation. 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.[11] 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.[12]

We follow four cohorts of first-grade children who entered the study schools in 2013, 2014, 2015, and 2016, comprising a total of 28,533 students with at least one test in either local language, English or math (Appendix Table A2, Panel A). Panel A of Appendix Table A3 describes the sample of students in the study, which varies both in the number of students sampled each year, and whether students were sampled at baseline or endline.[13] Students

subset of the schools in the reduced-cost version of the program

[10] 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.

[11] 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.

[12] 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.

[13] 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 from each of the 128 schools—sampled either at baseline or endline (Cohort 2). 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. 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.

were sampled stratified by gender and classroom. Across the five years of the study, a total of 1,480 teachers taught our sampled students, with approximately two teachers per grade (Appendix Table A2).

3.2.3 Data

We use three types of data: measures of student learning, characteristics of students and teachers, and school records.

Student learning outcomes consist of test scores in local language reading, English reading, and Math. 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. This results in Cohort 4 students only being assessed in one year, when they were in grade one in 2016.

Reading ability is measured using the Early Grade Reading Assessment (EGRA), internationally recognized assessments 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. Both versions of the EGRA that we use include 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 took an oral test to measure English speaking and listening ability instead. We measure math ability based on questions that measure numerical pattern recognition, one- and two-digit addition and subtraction, and matching numbers to objects. Math tests were self-administered, led by facilitators in a group setting. For each subject, we construct indices by first standardizing the separate test components against the control group for each student-year-grade combination, 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. We then standardize each test score index against the control group separately for each year and grade.

Throughout our analysis we include student-level controls for age and gender. We also use teacher characteristics from teacher surveys and employee rosters.[14] From these surveys

[14] 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.

and rosters, we have information on each teacher's age, gender, years of experience teaching, years as well as level of education.[15]

Appendix Table A4 presents descriptive statistics for students and teachers, 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, 40-45 percent female, with 15 years of education and 14 years of experience. Most teachers have a Certificate which is the minimum requirement for teaching in primary school and around 30 percent have qualifications above that (Panel B). We do not see any notable differences across study arms.

To shed light on sorting of teachers to grades we in addition make use of school records that document which teachers are teaching which grades. Here we have records for all grades one to seven.

3.2.4 Random Assignment and Balance

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.[16] 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. Schools are generally balanced across study arms in terms of student characteristics—age and gender—and teacher characteristics (see Appendix Table A4).

3.2.5 Attrition of Students and Teachers

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 in higher grades yet younger when looking within grade levels; otherwise we do not see any concerning differences in student

[15] When necessary, we convert all time-varying variables (i.e. age and experience) to their 2015 levels.

[16] 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.

attrition across study arms.[17] Student attrition might also depend on the teachers that they are exposed to. Appendix Table A6 presents the correlation between student attrition and teacher characteristics and shows that students with a female teacher are more likely to attrit in the control group but not in the reduced- and full-cost NULP study arms.

Of the 312 teachers we observe in 2014, 50 percent are still present in the sample in 2016—this differs somewhat between the control (40 percent), reduced-cost (48 percent), and full-cost treatment arms (62 percent).[18] Each year new teachers enter the sample and the likelihood of them staying the next year is around 60 percent. Appendix Table A7 presents the correlation between teacher characteristics and teacher attrition.[19] Overall we do not see many differences across characteristics; female teachers in the reduced-cost are somewhat more likely to attrit, while more educated teachers are less likely to attrit. Besides attrition, selection into the sample could also pose a problem as new teachers are coming into the sample each year. Appendix Table A8 presents the correlation between teacher characteristics and being an incoming teacher. Teachers who enter the sample are less likely to be female, less experienced and more educated, however this does not differ by treatment arm.[20]

4 Sorting and Analytical Samples

To motivate and understand our analytical approach and empirical strategy, we first discuss the sorting of teachers to schools, grades, and students. We then describe how we construct our analytical samples of students and teachers.

4.1 Sorting of Teachers and Students

Sorting of teachers to students has been extensively discussed in the literature as a potential threat to the estimation of teacher and classroom value-added (Rothstein 2017). Yet little is known about teacher and student sorting in developing-country schools, particularly in sub-Saharan Africa. In our data, there are three types of endogenous sorting to address: teachers

[17] 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.

[18] We compare 2014 to 2016 as 2014 was the first year with the full sample of schools and 2016 was the last year that collected data from students and teachers in grade one.

[19] Teacher attrition is defined as teachers only being observed once in our sample. We treat new teachers coming into the sample in 2017 as non-attritors as that is the last year of data collection and thus impossible to observe these teachers more than once.

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

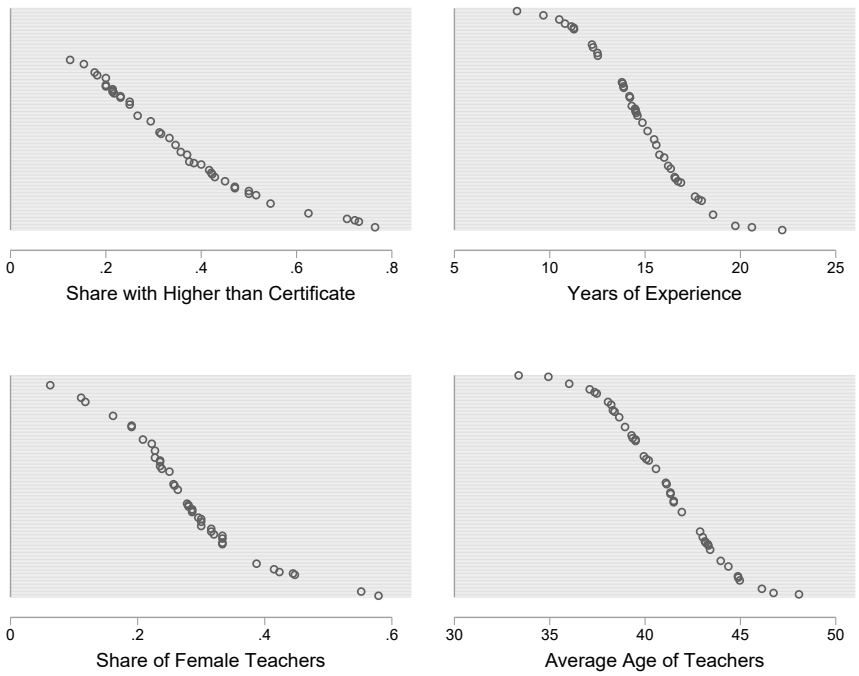
to schools, teachers to grades (within schools), and teachers to students (i.e., students to classrooms within grades).

Using data from teacher surveys and school records from the control schools we first describe the extent of non-random sorting by comparing teacher characteristics (i.e., the average rate of teachers having a higher level of education than a teaching certificate (the minimum requirement), years of experience, an indicator of being female, and age), across schools and grades. We then compare student baseline tests scores across classrooms.

4.1.1 Sorting of teachers to schools

Figure 5 plots the variation in teacher characteristics for the 42 control schools in our sample. The top left panel shows the share of teachers within a school, with the percent with a higher education level than a teaching certificate varying from 12 to almost 80 percent. The bottom left panel shows the share of female teachers within a school, varying from 6 to almost 60 percent. There is also substantial variation in number of years of teaching experience – from an average of 7 to 23 years. Together these results show substantial variation and that teachers are not equally distributed across schools.

Figure 5
Teacher Characteristics by School

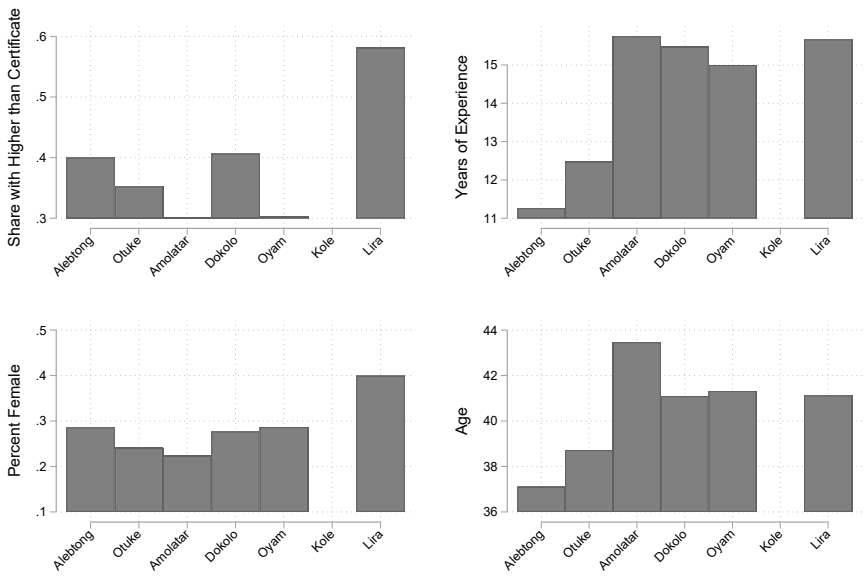


Notes: The panels in this figure graph the distributions of average teacher characteristics estimated within each of the 42 control schools in our sample (averaged within schools and across years).

To examine whether teachers with similar characteristics systematically sort to certain schools, we group schools into their geographical district (with seven districts in our sample), and plot the average of each teacher characteristic, sorting by district-level GDP. Figure 6 presents these results. Lira (graphed on the far right), contains the region's capital and is the most wealthy district in our sample. Lira also has teachers with the highest qualifications, the highest share of female teachers and also high levels of teacher experience. Teachers in schools in the poorest district, Alebtong (graphed on the far left), are younger and have lowest levels of experience.[21]

[21] Schools with low average levels of years of experience may reflect inability to either recruit experienced teachers, and/or retain teachers.

Figure 6
Teacher Characteristics by District



Notes: The panels in this figure graph the distributions of average teacher characteristics (calculated from the 42 control schools) across seven districts in the Northern Uganda region. The districts are sorted by district-level GDP (from poorest to richest).

These figures suggest that higher-skilled teachers sort into schools located in areas with more resources; our estimates of teacher value-added measure the within-school variance to address this.

4.1.2 Sorting of teachers to grades

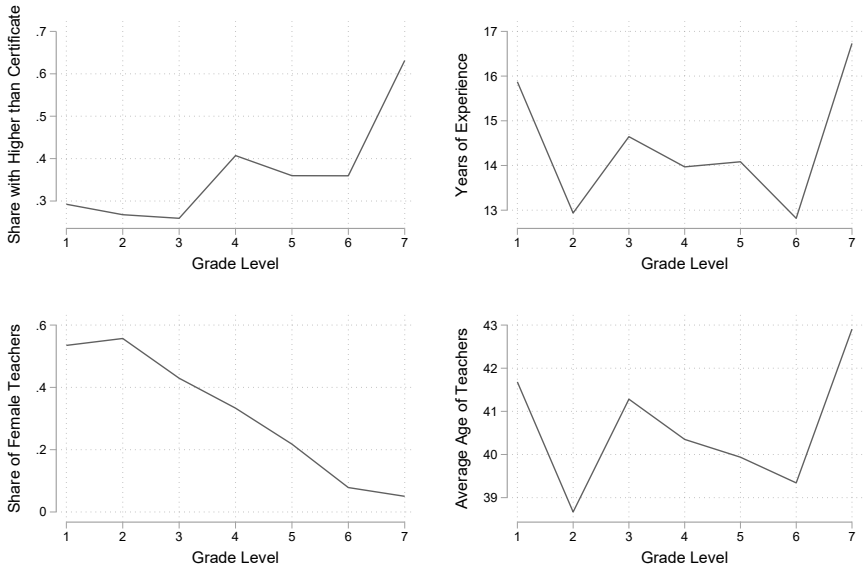
In Ugandan government primary schools, one teacher is typically assigned to one classroom, with multiple teachers teaching within a grade.[22] Within a school, head teachers have discretion to assign teachers to specific grades. One potential source of bias that could arise when estimating teacher value-added among students in multiple grades is if certain types of teachers are placed systematically in specific grades.

Figure 7 shows that teachers in higher grades – especially in grades 6 and 7 – are more likely to have higher qualifications than a teaching certificate and less likely to be female. With regard to experience and age we see a u-shaped relationship in which more experienced (and older) teachers are more likely to teach in grades 1 and 2 as well as in grades 6 and 7.

Anecdotally, because of the importance of the grade seven primary leaving exam, more emphasis from schools and parents is placed on higher grades, which could explain why

[22] In our sample, 88 percent of teachers teach in the same grade two years in a row.

Figure 7
Teacher Characteristics within School across Grade Levels



Notes: The panels in this figure graph the average teacher characteristics by grade level within each of the 42 control schools. We use data from all years from 2013 to 2017 and have approx. 300 to 400 teacher-year observations for each grade level.

teachers placed in higher grades have more education and teaching experience. Another explanation is that teachers with higher seniority may request to be placed in higher grades if those children are easier to teach.

Given that our data contain students and teachers across multiple grades and we find evidence of systematic sorting of teachers to grades, it is important to estimate the variation of value-added both within school and within grade.

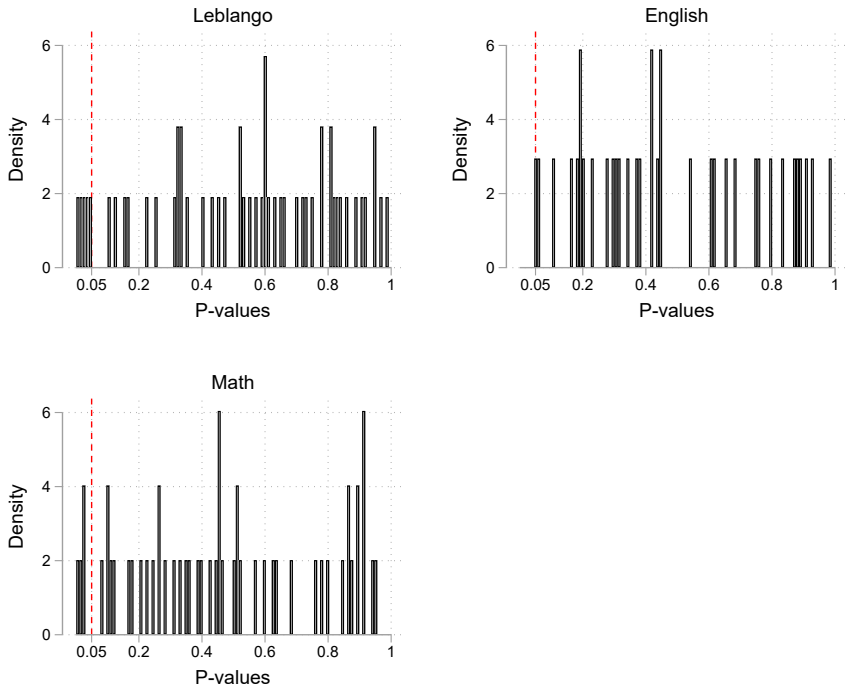
4.1.3 Sorting of students to classrooms within grades

A third potential source of bias in estimating value-added is non-random sorting of students to classrooms. In the United States, higher ability students tend to be more likely to receive instruction from higher ability teachers.

To investigate the degree of sorting of students to teachers/classrooms, we calculate the difference in prior year student test scores, between classes within-schools and grades each year, to test whether similar types of students sort into the same classroom in business-as-usual years (Horvath 2015). We find little evidence of student sorting: Figure 8 presents the p-values of the above mentioned tests indicating no sign of bunching below the significance level of 5%. In local language reading, 5 out of 53 comparisons (corresponding to 9%), show

significant differences in baseline test scores at the 5 percent level.

Figure 8
Differences in baseline test scores between classrooms



Notes: The panels in this figure graph the p-values of testing differences in average baseline test scores between classrooms within grades and schools within each of the 42 control schools. We use data from years with business-as-usual assignment only (2014 and 2015). The red vertical line mark a p-value of 0.05.

All in all, we find evidence of sorting of teachers across schools and grades but not across classrooms within grades.

4.2 Construction of Analytical Samples

In this section we describe our main analytical samples to estimate classroom and teacher value-added.

4.2.1 Annual Student Learning Gains

Our empirical strategy involves measuring the average gain in student learning attributable to a teacher in a given school year. Appendix Table A9 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. For every student-year observation with an endline test score in a given subject, we identify prior performance in

that subject. 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.[23] Because first grade students were not tested in English reading, we estimate English reading value-added only for students in grades two and above.[24] For students in grade two, we use oral English scores from the previous year to construct learning gains while for students in grades three, four, and five, we use their previous year English reading score (See Appendix Table A9).

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 an assessment. In that case, we impute students' missing prior test score as zero.[25]

4.2.2 Matching Students to Teachers

We match students to specific teachers using classroom registers and student reports. Across 58,774 total student-year observations for which we have at least one endline test score in local language, we are able to match approximately 99 percent to a teacher (see Appendix Table A3, Panels B and C).[26]

To limit estimation error due to sampling variation, we drop student-year observations with fewer than five students per teacher in a given year.[27] This removes 2,442 student-year observations, or 4.2 percent of the overall sample, bringing us to 55,702 student-year observations in local-language reading.

[23] This is motivated by the fact that 1) we only have baseline tests for grade one students in local language reading, only in 2013 and 2014, and even in these years we only have baseline tests for a subset of students who were sampled at baseline and 2) among the grade one 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 and we show our results under these alternative specifications for robustness.

[24] 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.

[25] For both local-language and English reading, around 10,000 student-year observations have missing prior test scores (corresponding to approx. 17% for Leblango and math and 27% for English). This does not vary between treatment arms.

[26] This rate does not vary systematically across year or treatment arm (99 percent in the full-cost treatment, 98 percent in the reduced-cost treatment, and 99 percent in the control). The most common reasons for not being able to match students to teachers include missing or misreported teacher names. 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.

[27] The rate of observations with fewer than 5 students per teacher does not vary much across randomization years or across school treatment arms (3.4 percent in full-cost, 4.2 percent in reduced-cost, and 5.1 percent in control schools; the p-value from an F-test testing equality across treatment arms is 0.12).

4.2.3 Analytical Samples

To address sorting across schools and grades, we need at least two teachers in each school (or grade) to purge out school (or grade) 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 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 would drop relatively more teachers from earlier years as we have more schools with only one teacher. To avoid this, we purge overall school effects instead of year-specific school effects. Similarly, many grades only have one classroom per year, thus to purge grade effects we require there to be at least two teachers per grade across all years (rather than within each year separately) allowing us to purge overall grade effects from the classroom value-added estimates. Table 1 Columns 1 and 2 shows the number of schools and teachers meeting these criterion in control schools, which represents the status quo in our setting.

Table 1
Analytical Samples for Control Schools

	Classroom Effects		Teacher Effects	
	Purging School Effects	Purging Grade Effects	Purging School Effects	Purging Grade Effects
	(1)	(2)	(3)	(4)
Schools	42	42	42	39
Teachers	361	322	152	124
– with characteristics	319	281	148	120
Classrooms	571	491	362	293
Pupils	8,814	7,940	7,624	6,260
Student-year obs	17,571	14,202	11,673	8,784

Notes: The 42 control schools were sampled in two phases: 12 in 2013 and an additional 30 in 2014.

To separate teacher effects from classroom effects, we need to observe a teacher over multiple years. We observe roughly 40 percent of the teachers teaching at least two years, which is the sample we use to calculate teacher effects, see Table 1, Columns 3 and 4. For both classroom and teacher effect estimates, we also perform analysis on a subset of teachers for whom we also have data on their background characteristics—roughly 80 percent of the teachers teaching in two years for classroom effects and 95 percent of these teachers for teacher effects. We use these teachers to estimate the correlation between teacher effectiveness and teacher characteristics.

5 Teacher Effectiveness under the Status Quo

5.1 Estimation Strategy

This section describes our empirical approach to estimating classroom and teacher value-added.

5.1.1 Classroom Effects

We begin by estimating classroom effects using the following "lagged-score" value-added model, separately for local-language reading, English reading, and math which takes prior student achievement into account to control for variation in initial conditions and treats the arguments in Equation (1) as additively separable (see e.g. Rivkin, Hanushek, and Kain 2005; Todd and Wolpin 2003):[28]

$$\begin{aligned} Y^k_{ijgs,t} =& \beta^k_1 Y_{ijgs,t-1} + \beta^k_2 Z_{ijgs,t-1} + \beta^k_3 X_{ijgs,t} + \lambda^k_{jgs,t} + \zeta^k_g + \beta^k_4 D_{ijgs,t} + \beta^k_5 ST_{ijgs,t} + \\ & \beta^k_6 Y_{ijgs,t-1}\zeta_g + \beta^k_7 Z_{ijgs,t-1}\zeta_g + \epsilon^k_{ijgs,t} \end{aligned} \tag{4}$$

where $Y^k_{ijgs,t}$ is the endline test score for subject k (Leblango, English or math) for child i taught by teacher j, in grade g, in school s, in year t. $Y^k_{ijgs,t-1}$ is the student's prior test score for the test of interest.[29] $Z_{ijgs,t-1}$ is a vector of prior scores for the other two assessments. Both of these capture previous family, school and unobserved individual factors as well as genetic endowments. $X_{ijgs,t}$ is a vector of individual characteristics, specifically gender and age. We include (expected) grade-level (ζ^k_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 $D_{ijgs,t}$.[30] Moreover, we include an indicator for the sample type $ST_{ijgs,t}$, 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^k_6 and β^k_7).

Our coefficients of interest are $\lambda^k_{jgs,t}$, which are classroom fixed effects (i.e., the effect of having a specific teacher in a specific year). These are estimates of the increase in learning attributable to being in a specific classroom in year t, and correspond to $C(M_{sjkt}, L_{sjkt}, E_{sjkt})$

[28] In a simulation exercise, Guarino et al. (2015) find, that the "lagged-score" model performs best in most scenarios. We perform robustness checks to the choice of model below.

[29] For grade 1 these are all set to zero in our main specification. For grades 2 and above this is prior end-of-year test scores.

[30] We perform additional robustness checks (described below) to address missing prior scores.

from Equation (1). To estimate a full set of classroom effects, we omit the constant term from the regression. Year fixed effects are implicit in the classroom fixed effects. We use use all possible observations to estimate $\lambda^k_{jgs,t}$. We estimate $\lambda^k_{jgs,t}$ on our entire analysis sample, prior to restricting the data to subsamples of interest such as teachers with data in multiple years.

When estimating Equation (4) we use both within- and between-school variation. This means that the estimate $\hat{\lambda}^k_{jgs,t}$ picks up both classroom effects as well as grade and school effects that co-vary with classroom effects. Since teachers were not randomized to schools nor grade-levels, some of the evident variation in our estimated classroom effects likely results from sorting of students across grades or schools. To overcome these issues we re-scale the classroom effects $\hat{\lambda}^k_{jgs,t}$ to be relative to the school or school-by-grade mean of the estimated classroom effects and thereby only consider the within-school/within-school-grade variation in the classroom effects (Araujo et al. 2016):

$$\textit{within-school}: \qquad \hat{\gamma}^k_{jgst} = \hat{\lambda}^k_{jgst} - \frac{\sum_{c=1}^{C_s} N_{cs}\hat{\lambda}^k_{jgst}}{\sum_{c=1}^{C_s} N_{cs}} \tag{5}$$

$$\textit{within-grade}: \qquad \hat{\gamma}^k_{jgst} = \hat{\lambda}^k_{jgst} - \frac{\sum_{c=1}^{C_{sg}} N_{cs}\hat{\lambda}^k_{jgst}}{\sum_{c=1}^{C_{sg}} N_{cs}} \tag{6}$$

where C_s is the number of grade one to four classrooms in school s, C_{sg} is the number of classrooms in grade g in school s and N_{cs} is the number of sampled students in classroom c in school s, and $\hat{\lambda}^k_{jgs,t}$ is the estimated classroom effect for a specific classroom. This approach nets out (in expectation) all school-level/school-by-grade-level factors and thereby provides a lower bound on the degree of variation in the classroom effects, since some of the across-school and across-grade variation in classroom effects represents real differences in teaching quality.

5.1.2 Teacher Effects

The estimated classroom effects from Equations (4), (5) and (6) 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. We estimate teacher effects using the demeaned classroom effects with the following equation:

$$\hat{\gamma}^k_{jgs,t} = \delta^k_{jgs} + \omega^k_{jgs,t} \tag{7}$$

where δ^k_{jgs} is a vector of teacher indicators and can be interpreted as the "permanent" component of teacher effectiveness. With this approach, we assume that all time variation in the classroom effects is due to transitory shocks and not changes in actual teacher effectiveness.

5.1.3 The Variance of Classroom and Teacher Value-added

The variation in classroom and teacher effects can be interpreted as the extent to which the classroom or teacher a student is assigned to matters for learning outcomes. A small variance means that classrooms or teachers are very similar and thus it does not matter which classroom or teacher a student gets. Conversely, a large variance means that classrooms or teachers are very different and thus it matters a great deal which classroom or teacher a student is assigned to. Before making this interpretation we need to adjust the estimated variance to account for sampling variation, due to estimating classroom and teachers with finite samples of students. The smaller the number of students per classroom or teacher, the more likely that the estimated total variance of value-added will be high or low due to random chance. To address this issue, we follow the approach suggested by Araujo et al. (2016).[31] 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:[32]

$$\hat{V}_{\text{corrected}}(\hat{\gamma}^k_{jgs,t}) = V(\hat{\gamma}^k_{jgs,t}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{[(\sum_{c=1}^{C_s} N_{cs}) - N_{cs}]}{N_{cs}(\sum_{c=1}^{C_s} N_{cs})}\hat{\sigma}^2\right\} \tag{8}$$

where $\hat{\sigma}^2$ is the variance of the estimated residuals, $\hat{\epsilon}^k_{ijgs,t}$, from Equation (4). C is the overall number of classrooms in the sample, and N_{cs} is the number of students in classroom c in school s.

When we use only within-grade variation the correction changes slightly to:

$$\hat{V}_{\text{corrected}}(\hat{\gamma}^k_{jgs,t}) = V(\hat{\gamma}^k_{jgs,t}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{[(\sum_{c=1}^{C_{sg}} N_{cs}) - N_{cs}]}{N_{cs}(\sum_{c=1}^{C_{sg}} N_{cs})}\hat{\sigma}^2\right\} \tag{9}$$

where we sum to the school-by-grade level (C_{sg}) instead of the overall school-level (C_s).

[31] The procedure is analogous to an Empirical Bayes approach. The difference is that the procedure we use 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.

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

$\hat{V}_{\text{corrected}}(\hat{\gamma}^k_{jgs,t})$ is our measure of interest when discussing the distribution of classroom effects. We correct the variance of the teacher effects for sampling variation in the same manner.[33]

5.1.4 Verifying Random Assignment of Students to Classrooms

Even after purging out school and grade-level effects, endogenous sorting of students to teachers *within* schools and grades can introduce bias to value-added estimates (e.g. Chetty, Friedman, and Rockoff 2014; Rothstein 2010; Goldhaber and Chaplin 2015; Kinsler 2012). To explore the severity of this potential bias, we utilize the random assignment of students to classrooms within grade levels in three of the five years of the study.

In 2013, 2016, and 2017, we explicitly instructed head teachers to randomly assign students to teachers/classrooms within grade levels (Appendix Table A1, Panel B).[34] Accordingly, if randomization was successful then value-added estimates obtained from random assignment years should by definition be free of sorting. To assess the degree of compliance with the random assignment of students to classes in 2013, 2016, and 2017 we test if teacher characteristics are orthogonal to baseline student characteristics. Appendix Table A10 presents regressions of baseline student characteristics on teacher characteristics. While there are a few statistically significant coefficients—more educated teachers are less likely to have older students (likely to be repeaters) and more likely to have students with higher English skills—the majority are small and insignificant.

As a second check for balance across randomly assigned students to teachers, we use the same method as in Section 4.1.3 using random assignment years instead of business-as-usual years (Horvath 2015). Appendix Figure A1 presents the distribution of p-values from regressing baseline test scores on classroom dummies within each year, school, and grade-level. For local-language baseline test scores, we find that out of 90 tests (within school, year and grade-level) 8 yield a significant difference (at the 5 percent level), corresponding to 9 percent. For English and math the percent is 5 and 14, respectively.[35] In our analysis,

[33] For the correction of the variance of the teacher effects we use the following adjusted form of Equation (8): $\hat{V}_{\text{corrected}}(\hat{\theta}^k_{jgs}) = V(\hat{\theta}^k_{jgs}) - \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{[(\sum_{t=1}^{T_s} N_{ts}) - N_{ts}]}{N_{ts}(\sum_{t=1}^{T_s} N_{ts})}\hat{\sigma}^2\right\}$, where $\hat{\sigma}^2$ is the variance of the residuals, $\hat{\epsilon}^k_{ijgst}$, from Equation (4). 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, $\hat{V}_{\text{corrected}}(\hat{\theta}^k_{jgs})$ is our measure of interest when discussing the distribution of teacher effects.

[34] 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.

[35] Statistically significant differences occur mostly in classrooms grade 3 and above, but across different schools— only one school has more than one grade-level in which we cannot reject sorting.

we compare the results using data from all years of the study with the results estimated separately in years where teachers were randomly assigned to classrooms.

5.2 Results

In this section we present estimates of classroom and teacher effects under the status quo, focusing on control-group schools only. In addition, we provide various robustness and sensitivity analyses including comparing our preferred value-added estimates to those calculated from years with random student-to-teacher assignment.

5.2.1 Variation in Classroom and Teacher Value-Added

Table 2 presents our estimates of teacher and classroom value-added, measured in terms of standard deviations of student performance on the end-of-year assessments. We present our estimates with corrections for sampling variance and present school-level cluster-bootstrapped confidence intervals in square brackets.[36]

Panel A of Table 2 shows the results that include both between- and within-school variation to estimate classroom and teacher effects. Columns 1 and 2 present the results for local language reading. After correcting for sampling variation, a one-SD increase in classroom quality increases student performance in local-language reading by 0.36 SDs; for teacher effects, the estimate is 0.27 SDs (Panel A, Columns 1 and 2). Columns 3 and 4, as well as 5 and 6 present corresponding results for English and math, respectively. Here, the estimates for classrooms including school effects are somewhat larger at 0.52 SDs for English and 0.51 SDs for math; the teacher effects are also larger at 0.43 and 0.42 SDs respectively. Because the estimates in Panel A 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 and grades. By implication, these estimates are upper bounds on the variance of the true λ_{jgst} (classroom effects) and δ_{jgs} (teacher effects). Teacher effects are between 17 and 25 percent smaller than classroom effects.

In Panel B 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. The estimated variance of teaching quality for local-language reading is slightly smaller, with the estimate showing that a one SD increase in classroom quality is associated with an increase in student performance of 0.33 SDs, and a one SD increase in teacher quality is associated with an increase of test scores of 0.24 SDs. The

[36] A full set of results including uncorrected results are presented in Appendix Table A11.

Table 2
Classroom and Teacher Value-Added: Control Schools

	Leblango		English		Math	
	Classroom	**Teacher**	**Classroom**	**Teacher**	**Classroom**	**Teacher**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Including School Effects						
Corrected SD	0.36	0.27	0.52	0.43	0.51	0.42
	[0.22,0.50]	[0.16,0.38]	[0.37,0.67]	[0.30,0.56]	[0.39,0.63]	[0.33,0.51]
Observations	17,571	11,673	10,865	6,333	17,422	11,568
Panel B: School Effects Purged						
Corrected SD	0.33	0.24	0.31	0.22	0.41	0.30
	[0.19,0.47]	[0.14,0.34]	[0.22,0.41]	[0.13,0.31]	[0.29,0.53]	[0.21,0.39]
Observations	17,571	11,673	10,865	6,333	17,422	11,568
Panel C: School-by-grade Effects Purged						
Corrected SD	0.11	0.09	0.17	0.11	0.30	0.18
	[0.06,0.16]	[0.07,0.11]	[0.12,0.22]	[0.07,0.15]	[0.24,0.36]	[0.15,0.21]
Observations	14,202	8,784	8,757	4,777	14,073	8,698
Panel D: Random Assignment Years - School-by-grade Effects Purged						
Corrected SD	0.12	0.09	0.15	0.12	0.24	0.18
	[0.06,0.18]	[0.05,0.13]	[0.11,0.19]	[0.09,0.15]	[0.21,0.27]	[0.15,0.21]
Observations	5,963	2,315	4,647	1,672	5,915	2,289

Notes: 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Control schools (N=42) did not receive the NULP intervention.

variation in classroom and teacher effects for English is very similar to the estimates for local language, while the estimates for math are somewhat larger.

To put the differences between Panel A and Panel B 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, then the estimates without school effects measure the true variance of teacher value-added in the entire population of teachers. If teachers perfectly sort to schools such that all of the most-effective teachers work together in one school, with the least effective in another school, 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. Accordingly, our results suggest that sorting of teachers to schools is primarily affecting teacher effectiveness in English and math.

As documented in Section 4, there is also sorting of teachers to grades within schools. Accordingly, we further limit our analysis to only use within-grade between teacher variation in Panel C. Doing this we see a rather large reduction (of roughly two-thirds) in both classroom and teacher effects for the Leblango estimates. We also see reductions in the estimates for English and math.[37] While the estimates in Panel C have the advantage of reducing bias due to non-random sorting across grade levels, we lose some sample size because we have fewer teachers per grade than per school.

Lastly, in Panel D, we restrict the sample to years where teachers were randomly assigned to classrooms. The change from panel C is negligible which is in line with the results in Appendix Figure A1. Taken together, our preferred estimates are those in Panel C, as we take into account bias from non-random sorting of teachers to students yet keep the largest possible sample. In Table 2 we only present results with the corrected SDs. However, in Figure 9, we show that the correction reduces the estimated SD of teacher effects by about 10 percent for most of our estimates. The SD correction only creates a statistically significant difference when we purge school-by-grade effects but do not isolate the stable component of teacher effects.

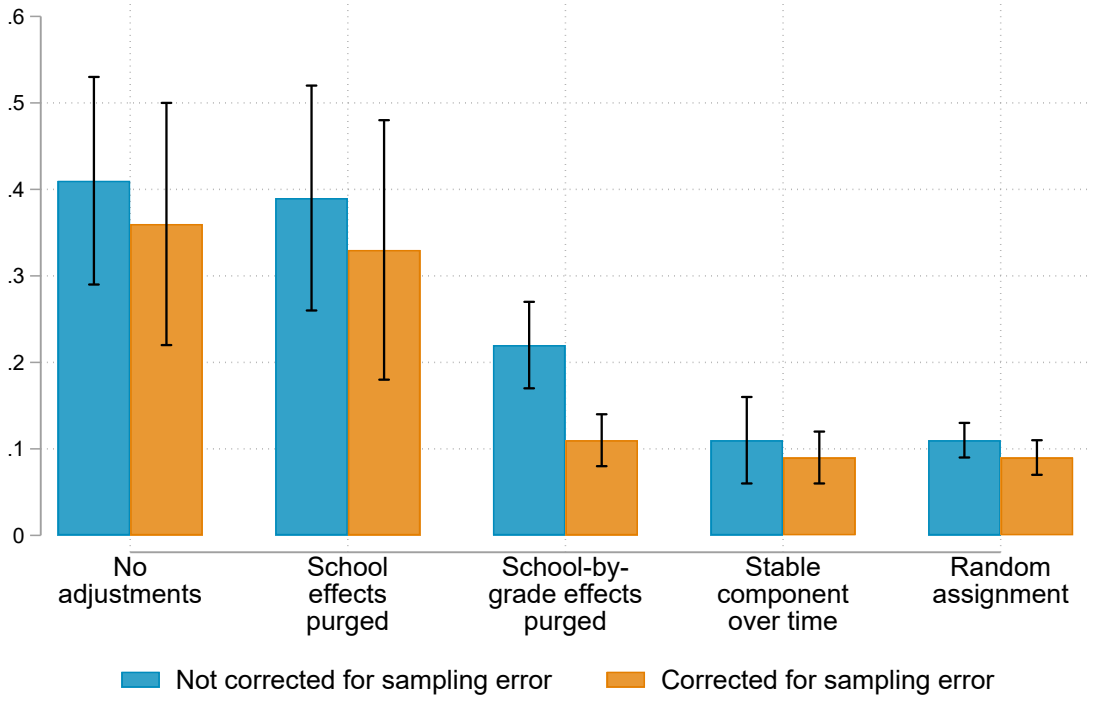
As mentioned above, classroom effects contains both the true effectiveness of a teacher as well as other time-varying classroom effects such as peer dynamics or conditions on the day of testing. Teacher effects, on the other hand, purge any year-on-year fluctuations in classroom effects. The difference in our estimated classroom and teacher effects gives an indication of the fluctuation across years in our context. For Leblango, the year-to-year fluctuations are roughly two-thirds of the teacher effects while they are larger than the teacher effects for both English and math. This suggests that the year-specific classroom shocks are very important in this context.[38]

Teacher value-added in Leblango reading is positively correlated with English—the estimates for the two subjects (after purging grade effects) have a correlation coefficient of 0.53. 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). Although weaker, teacher value-added for math is also positively correlated with the two language subjects with a correlation coefficient of 0.36 with English reading, and 0.l6 with Leblango reading. This suggests that teachers are relatively broad in their effectiveness—effective teachers in one subject are also likely to be effective teachers in other subjects.

[37] Appendix Table A12 shows that the reduction in estimates after purging school-by-grade effects is solely do to the change in variation and not the change in sample.

[38] We calculate the SD of the year-specific classroom shocks as follows: Leblango: $0.06 = \sqrt{0.11^2 - 0.09^2}$; English: $0.13 = \sqrt{0.17^2 - 0.11^2}$; and math: $0.24 = \sqrt{0.30^2 - 0.18^2}$

Figure 9
Classroom and Teacher Value-Added - Leblango Reading



Notes: This figure shows the estimated classroom and teacher value-added in Leblango using estimates from Table 2.

5.2.2 Sensitivity Analyses

We present several robustness checks for our main estimates of value-added in Table 2 that 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 student covariates are missing, or when we are missing baseline or prior-year test scores, and d) the use of alternative outcome measures.

Recall that only around 40 percent of the teachers used to calculate classroom effects were present in multiple years and thus there is a different sample of teachers used to calculate teacher effects. To test the potential effect of the difference in the two samples, Appendix Table A12 Panel A presents the equivalent classroom effects estimates using the same sample as the sample used to estimate teacher effects. The results are similar to those in Table 2. Relatedly, we lose observations when moving from purging school effects to purging school-by-grade effects. Appendix Table A12 Panel B presents the results when purging school effects using the same sample of teachers and students as in Panel C of Table 2. Again the results are similar to those in Panel C of Table 2 suggesting that the change in estimates

from Panel C to Panel D in Table 2 is a result of the estimation approach and not the change in sample.

Our preferred estimates in Table 2 condition on observations with at 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 fewer students on our results by re-estimating our results from Table 2, omitting teachers with less than 10 or 15 students. Appendix Table A13 shows that the estimates barely change among this sample.

We next address the fact that we impute missing student covariates—age or gender, or prior test scores. Appendix Table A14, Panel A presents the estimates without imputing the covariates, thus omitting 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 enrolled in 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. Panel B in Appendix Table A14 present results where we instead remove all first-grade students from the estimates. The variances of the classroom and teacher effects are only slightly affected relative to those in Table 2.

Finally, we present sensitivity to using alternative outcomes. Appendix Table A15 Panel A shows the results for 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 (4) with $\Delta Y^k_{ijgs,t} = Y^k_{ijgs,t} - Y^k_{ijgs,t-1}$. The results are similar to those in Table 2. In our preferred results we combine the test score components using the first component of PCA. As robustness we in Table A15 Panel B present the result using an alternative way of combining the test score components; here we simply calculate the mean of the test components for each student-year-grade observation and use that as the combined index. The choice of index barely change the results.

6 Treatment Effects on Teacher Effectiveness

In this section we present the causal effect of the NULP on the distribution of classroom and teacher value-added. Because the intervention focuses predominantly on local-language learning, we limit the results in this section to Leblango reading outcomes. After comparing the distribution of value-added across NULP treatment arms, we then turn to testing for rank preservation, as well as examining which teachers are associated with higher value-added or differences in value-added across treatment arms.

6.1 Framework and Estimation

The NULP is a highly effective educational intervention, increasing average student learning among students directly exposed to the program by 1.2 SDs in the full-cost version and 0.7 in the reduced-cost version for the main cohort of children after three years of exposure to the program (Appendix Table A16).[39] The NULP was given to different grades of students over time during our intervention (see Appendix Table A1). To estimate the effect of the NULP on teacher value-added, we use data from all cohorts of students in the full- or reduced-cost treatment arms, regardless of whether they were directly or indirect exposure to the treatment. Including students who were indirectly exposed across all cohorts of students in our data yields an overall average treatment effect of 0.54 SDs in the full-cost version and 0.22 SDs in the reduced-cost version—smaller, yet still sizable effects (see Appendix Table A16).

The NULP could have affected learning through either increased inputs or increased productivity/effectiveness of teachers, resulting in movements along the learning production functions outlined in our framework in Section 2. Given the large learning gains that we observe from the NULP, and the fact that generally returns to educational inputs alone have been found to be quite low, it is likely that the NULP increased teacher effectiveness, at least among some teachers.

To estimate the effects of the NULP on the distribution of value-added estimates, we calculate $\hat{V}_{\text{corrected}}(\hat{\gamma}_{cgs,t})$ and $\hat{V}_{\text{corrected}}(\hat{\delta}_{cgs})$, separately, for each of the three treatment arms. As described above, the NULP rolled out in treatment schools to different grades across years (Appendix Table A1). This means that teachers or students in some years or grades did not directly receive the NULP.[40] Our preferred estimates pool cohorts that would have been directly exposed to the NULP with those who were only indirectly exposed as the program rolled out; we provide sensitivity analyses for grades and years for which students/teachers would have directly received the NULP in a given year.

6.1.1 Rank preservation

Our framework shows how increasing learning inputs or training and supporting teachers can shift teacher effectiveness to produce higher learning outcomes, and that whether the distribution of effectiveness increases, decreases, or stays the same, depends on which teachers

[39] The estimated effects of the program are slightly different than those in Buhl-Wiggers et al. (2022) because that other paper restricts the sample to students with non-missing baseline covariates.

[40] Teachers/classrooms not directly exposed to the NULP in treatment schools would been in classrooms with the NULP materials (e.g., slates, readers, primers) from previous years, and for teachers that did not transfer grades or schools would have received NULP training and support in prior years.

are affected by the intervention. Testing for rank preservation in teacher quality provides insight into the changes in the distribution of teacher value-added.

We follow Bitler, Gelbach, and Hoynes (2005) and Djebbari and Smith (2008) and test whether fixed teacher covariates have the same means in a given quantile of the teacher value-added distribution. The approach is as follows: First we divide our estimates of classroom and teacher effects into quantiles. Then we demean the teacher characteristics separately by treatment arm and compute means of each characteristic within each quantile of the teacher and classroom effects. Finally, we regress the quantile-specific means on indicators for quantile, study arm, and the interaction between the two. More formally, we estimate the following equation:

$$Z^l_{dq} = \alpha^l_0 + \alpha^l_1 FullCost_s Q_q + \alpha^l_2 ReducedCost_s Q_q + \alpha^l_3 Q_q + \zeta^l_{strata} + \varepsilon^l_{dq} \tag{10}$$

where, Z^l_{dq} is the mean of teacher characteristic l within each treatment arm d and quantile q. $FullCost_s$ is an indicator for whether the school received the full-cost program. $ReducedCost_s$ is an indicator for whether the school received the reduced-cost program. Q_q is an indicator for the quantile and ζ_{strata} is a set of stratification-cell fixed effects. α^l_1 and α^l_2 is our coefficients of interest and measure the differences in mean characteristics Z between the control and full-cost or reduced-cost program for each quantile of the teacher/classroom effects distribution. To estimate the full set of interactions $\alpha^l_1 FullCost_s Q_q$ and $\alpha^l_2 ReducedCost_s Q_q$ we omit the main effects $FullCost_s$ and $ReducedCost_s$ from Equation (10).

6.1.2 Correlation between value added and teacher characteristics

To see if certain teachers characteristics are correlated with value added, we estimate the following equation:

$$\hat{\delta}_{cgs} = \beta_0 + C'_{cgs}\beta_1 + \phi_s + \varepsilon_{cgs} \tag{11}$$

where $\hat{\delta}_{cgs}$ are our estimated classroom or teacher effects purged of school-by-grade effects, C_{cgs} is a vector of teacher characteristics that includes sex, age, years of experience, and level of education. ϕ_s indicates school-level fixed effects. We estimate this separately by study arm to understand what factors are the most important predictors of higher value-added, and compare the coefficients across study arms.

6.2 Results

We first present the effects of the NULP on the variation of classroom and teacher effects, then test for rank preservation. We then correlate our value-added estimates with teacher characteristics to understand which teachers are most effective, and lastly provide several robustness checks. All of our results in this section use learning outcomes in local-language reading, the focus of the NULP intervention.

6.2.1 Impact of NULP on the Distribution of Value-Added

In Table 3, we show how the NULP affects the variance of classroom- (Columns 1-3) and teacher-value added (Columns 4-6) estimates. Columns 1 and 4 show the results for teachers in control group schools, identical to the results in Table 2 Panel C, Columns 1 and 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 3
Heterogeneity of Value-Added by NULP Study Arm

	Classroom Effects			Teacher Effects		
	Control	Reduced-cost	Full-cost	Control	Reduced-cost	Full-cost
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD	0.11	0.22	0.30	0.09	0.14	0.16
	[0.06,0.16]	[0.17,0.28]	[0.25,0.35]	[0.07,0.11]	[0.11,0.17]	[0.13,0.19]
Observations	14,202	15,921	15,313	8,784	10,793	10,537
Classrooms/Teachers	491/322	544/340	502/306	293/124	345/141	334/138
Schools	42	43	42	39	41	37

Notes: All estimates are purged of school-by-grade effects by subtracting off a weighted school-by-grade 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 ([0.10;0.25]) and teacher ([0.04;0.11]) effects.

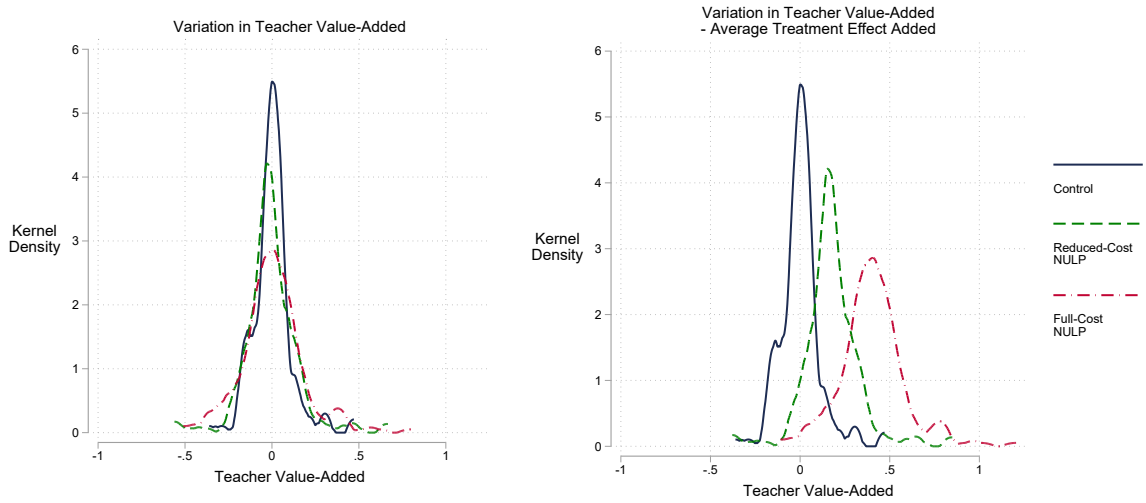
The NULP increases the variance of classroom and teacher effects. The corrected standard deviation of classroom effects in Leblango increases from 0.11 SDs in control schools to 0.22 SDs in reduced-cost and 0.30 SDs in full-cost program schools (Columns 1-3). The estimated increases in the standard deviation of teacher effects due to the program are somewhat smaller from 0.09 SDs in control schools to 0.14 in reduced-cost and 0.16 SDs in full-cost schools (Columns 4-6). To formally test the difference between the study arms we bootstrap the difference between arms and examine the fraction of re-samples for which the difference is zero or smaller.[41] 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.

Focusing solely on the control-group results, a 0.09 SD increase in teacher effectiveness suggests that moving the worst (10th-percentile) teachers to the level of the best (90th-percentile) is associated with an increase in student learning of 0.22 SDs. Yet, if the worst performing teachers are the most difficult to move, this interpretation would be overstating the actual gains. When actually providing inputs and support for teachers the average teacher increase their effectiveness by 0.40 SDs while the treatment effect is only about 0.10 SDs for the 10th percentile. This suggest that focusing on improving all teachers result in more learning on average as opposed to only focusing on moving the worst performing teachers.

In summary, two things happen in response to the NULP: 1) teachers, on average, are becoming more effective (level shifts in student learning); and 2) teachers become more heterogeneous in their ability to affect learning (increased variance of teacher effectiveness). Figure 10 illustrates these two effects. Our estimates in Table 3 present the within-school-grade estimates purging out any level differences across grades and schools—and thus also purging out the NULP treatment effect since the intervention took place at the school level. Accordingly, to illustrate both the shift in levels and in variation due to the NULP on the distribution of value-added, we present the graphs of the distribution of teacher value-added, manually adding the average NULP treatment effect on teacher effectiveness in full-cost and reduced-cost schools.

[41] Formally, we calculate the difference 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

Figure 10
Treatment Effects on Teacher Value-Added



Notes: This figure presents the distribution of teacher effects estimated separately by NULP treatment arm either simply purging grade and school effects (left-hand-side), or mechanically adding the average treatment effect; 0.18 for the reduced-cost and 0.40 for the full-cost program. These averages are estimated from regressing the non-demeaned teacher effects on treatment indicators.

6.2.2 Sensitivity Analyses

While the NULP intervention was only implemented for certain grade-levels in certain years (see Appendix Table A1), more than half (60 percent) of the teachers in our sample were provided with training at some point during the years of the intervention and data collection. Of these, 74 percent of teachers were treated one year and 26 percent were treated multiple years. We perform four related sensitivity tests in Appendix Tables A17, A18, A19 and A20.

First, because the NULP was only fully implemented between 2013 to 2016 (see Appendix Table A1). Appendix Table A17 shows our main estimates of value-added across NULP treatment arm, restricted only to the years 2013-2016 (ie. omitting data collected in 2017). Including 2017—when NULP was not directly implemented—could mute any differences across the treatment arms and control. Indeed, we see an increase in the difference across treatment arms in both classroom and teacher effects. This is mainly driven by a change in the control group estimates of the variance of the classroom effects. The estimates of classroom and teacher effects in the reduced-cost and full-cost treatment arms increase slightly.

Second, we restrict our sample to only include cohorts of students and teachers who would have been directly trained and supported by the NULP for each of the NULP treatment

groups, and the corresponding students and teachers in the control group. Here we include teachers in years that they directly received the NULP as well as all the subsequent years they appear in the data—this assumes that teachers can make use of the NULP tools in years after they received training (Appendix Table A18). Again, the variation of the classroom effects is somewhat smaller for the control group, otherwise these estimates show similar patterns to Table 3. A more restrictive approach would be to only consider a teacher treated in the year that they directly receive NULP training and support. Using this approach we can only estimate classroom effects because of the limited number of teachers that taught in higher grades with each subsequent year, thus being exposed to the NULP multiple times. For example, teachers would have to have taught grade 1 in 2013 or 2014, grade 2 in 2015 and grade 3 in 2016. Appendix Table A19 presents these results showing a similar pattern as Appendix Table A18.

Finally, a less restrictive way to run this would be to expand the sample to teachers who were trained multiple times across all years (ie., not just years that they were directly exposed to the NULP), to see how the results are affected. If teachers are only treated once and the effects from the training and support does not persist, then the NULP treatment effect would act as a year-specific shock, which would be purged out in the teacher effects. The results are presented in Appendix Table A20 and yield similar patterns to Table 3 although weaker effects in reduced-cost program schools.

These sensitivity checks do not affect our conclusion that the NULP increased the variance of classroom and teacher value-added.

6.2.3 Who are the Most Affected Teachers?

The finding that the NULP—a highly effective teacher-training program—increases the spread of teacher effectiveness means that some teachers improve more than others. If there is rank preservation such that the effects of the program is largest for the most effective teachers, we can interpret the results as a "skills-beget-skills" story, corresponding to Figure 3c in Section 2.3.

Rank preservation 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. An alternative story would be if instead changed rank as a consequence of the NULP—in this case, an increase of the variance of teacher value-added could only happen if the program was especially effective for the low-performing teachers and they "leap-frogged" the best-performing teachers, corresponding to Figure 4, a and b in Section 2.3. This seems intuitively unlikely but not something we can rule out theoretically.

Table 4 presents the results of tests for rank preservation where we focus on comparisons

between the full-cost program and control-group schools; the results from rank preservation tests are similar when we compare the reduced-cost program schools to control schools. Each column represents a fixed teacher background variable (age, gender, experience and degree obtained). Each row summarizes the difference in the average of each background characteristic between the full-cost and control schools, that are within each quartile of the distribution of classroom value-added (CVA) in Panel A or teacher value-added (TVA) in Panel B, estimated in Panel C, Table 2 for Leblango.

For each teacher characteristic, we test the null of zero difference in the means within the population quartile of value-added between the full-cost program and the control, a total of 4x4=16 tests. Under the (surely incorrect) assumption of independence of the different tests, we would expect about two or three rejections. Table 4 shows that 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.

There are three caveats to these results. First, we do not have characteristics for all teachers, so we cannot test rank preservation using the full sample of teachers. Second, the power of these tests is limited by the fact that teacher characteristics are only weakly correlated with teacher effects (see Table 5). 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.

Our last set of analyses to understand which teachers perform best, and which teachers are most affected by the NULP, correlates teacher variables with our measures of effectiveness. If we can predict effectiveness using teacher characteristics, educators could more successfully recruit and hire teachers who would be more likely to be successful in a classroom.[42] Similarly, if we know which teachers are most or least responsive to intervention like the NULP, we could target teacher training programs.

Using data on teacher gender, years of experience, age, and education level, we first describe how teacher characteristics correlate with our classroom and teacher value-added estimates in the control schools. Table 5, Panel A, Columns 1 and 4 show regressions of Classroom effects and Teacher effects, respectively, on teacher characteristics in control schools.

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

Table 4
Tests of Rank Preservation

	Age (1)	**Female** (2)	**Experience** (3)	**Above Certificate** (4)
Panel A: Classroom Effects				
First quartile of CVA	0.687	0.037	0.803	0.034
	[–2.317,2.167]	[–0.120,0.113]	[–2.063,2.040]	[–0.148,0.138]
Second quartile of CVA	–1.131	–0.062	–1.711	0.007
	[–2.173,2.018]	[–0.123,0.125]	[–1.983,1.938]	[–0.124,0.117]
Third quartile of CVA	0.196	–0.013	–0.348	–0.044
	[–1.890,1.817]	[–0.143,0.136]	[–1.903,1.740]	[–0.144,0.140]
Fourth quartile of CVA	–1.354	0.026	–0.498	0.003
	[–2.061,1.949]	[–0.126,0.127]	[–2.187,2.257]	[–0.107,0.101]
Observations	901	926	888	895
Panel B: Teacher Effects				
First quartile of TVA	–2.687	0.187	0.115	–0.017
	[–3.804,4.383]	[–0.228,0.245]	[–3.367,3.355]	[–0.213,0.228]
Second quartile of TVA	–0.152	–0.075	–1.849	–0.066
	[–3.334,3.350]	[–0.228,0.220]	[–2.982,2.876]	[–0.218,0.214]
Third quartile of TVA	–1.740	0.094	–0.651	–0.040
	[–3.573,3.094]	[–0.205,0.208]	[–3.297,3.607]	[–0.204,0.193]
Fourth quartile of TVA	0.339	–0.179	–1.118	0.064
	[–3.518,3.811]	[–0.197,0.222]	[–4.221,4.152]	[–0.210,0.199]
Observations	615	625	612	617

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 and TVA=Teacher Value Added.

We find some indication that teachers with higher education levels have lower classroom effects while teachers with more than three years of experience have higher classroom effects. In general, however, we find no stable patterns of predictors of classroom or teacher value-added. Columns 2 and 3, as well as 5 and 6, present the results separately for the reduced- and full-cost treatment schools to see if the correlation between teacher characteristics and teacher effectiveness varies with receiving the two NULP program versions. Again, there are no strong patterns.

For a sub-sample of teachers, we have additional data on self-reported days missed of school as well as their reading performance in Leblango measured with the same EGRA test conducted for their students. In Panel B we present the correlation between these and our classroom and teacher value-added and allow the relationship to vary by treatment arm. We find no strong patterns, suggesting that overall we cannot say much about who are the most affected teachers based on their characteristics.

Table 5

Teacher Value-Add Correlation with Teacher Characteristics

	Classroom Effects			Teacher Effects		
	Control	Reduced-Cost	Full-Cost	Control	Reduced-Cost	Full-Cost
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: All teachers with characteristics						
Above Certificate	-0.038*	-0.003	-0.021	-0.028	-0.003	0.037
	(0.019)	(0.036)	(0.082)	(0.039)	(0.053)	(0.085)
Female	0.027	0.003	0.030	0.022	-0.043	0.026
	(0.028)	(0.039)	(0.059)	(0.038)	(0.068)	(0.055)
$>$ 3 years of	0.082**	-0.008	-0.177	0.063	-0.007	-0.129
experience	(0.037)	(0.050)	(0.117)	(0.061)	(0.045)	(0.189)
Age	0.002	-0.001	-0.000	0.001	-0.002	0.000
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	485	550	518	120	137	135
R-squared	0.044	0.022	0.061	0.199	0.203	0.136
Panel B: Teachers with survey data and characteristics						
Days missed	0.006	-0.012	-0.016	0.064	0.000	-0.044
	(0.005)	(0.008)	(0.022)	(0.085)	(0.004)	(0.035)
Teacher test score	0.004	-0.003	-0.007	-0.010	-0.002	0.004
(Leblango)	(0.005)	(0.005)	(0.005)	(0.013)	(0.010)	(0.008)
Observations	101	112	108	39	38	42
R-squared	0.306	0.306	0.540	0.731	0.789	0.765
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors are clustered by school, in parentheses; $*p < 0.10, **p < 0.05, ***p < 0.01$. The dependent variables are school-by-grade demeaned teacher and classroom effects. All regressions include school fixed effects. In Panel B controls include: A dummy for having above certificate in education level, dummy for being female, dummy for having more than three years of experience and age.

7 Conclusion

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.

Using five years of data from students and teachers from northern Uganda, we find substantial variation in teacher effectiveness. This variation increases when teachers are randomly exposed to a highly-effective literacy program. Our findings have several implications for understanding how teachers contribute to the production of learning.

First, despite the overall low learning levels in our setting in rural Uganda, we find substantial variation in teacher effectiveness—in other words, some teachers are much more effective in increasing learning than others. We also show that in our setting, sorting of students to teacher by ability is not an important issue for estimation. Instead, we provide evidence descriptively and in our estimation of value-added, that sorting of teachers to schools and grades is an important issue. This has important implications for researchers using these methods with data that span across grades. More descriptive work could shed light on classroom dynamics and sorting of teachers and students in Africa.

Second, the NULP resulted in massive average gains in student learning. We find that the NULP also increases the variance of teacher effectiveness—most likely by making the most effective teachers even more effective. This implies that successful educational interventions might increase inequality in education as more skilled teachers are better able to make use of their training. One potential avenue for future research is to examine which interventions affect less-effective teachers.

The variance in teacher value-added is usually interpreted as the scope for improving learning outcomes through teachers. Yet simply comparing the variance in teacher effective-

ness across very different settings may not be informative. In low-income countries, even the best teacher might have scope to improve, and a low variance of teacher 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.

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.

Our approach—which combines 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 conjecture regarding the hypothetical effect of moving teachers up the distribution of quality, we can observe how the distribution actually shifts.

This paper presents the first estimates of the distribution of teacher effectiveness from sub-Saharan Africa. Our study relies on data from Northern Uganda, and so our results may not generalize to other parts of Africa. However, many features of our setting are common to African schools: enrollment is high, classes are large, learning levels are low, and teacher absence is common (Bold et al. 2017). Our results show that even with all of these constraints, there is substantial variation in teacher value-added and meaningful scope for improving teacher effectiveness. Future research should explore how the distribution of teacher value-added differs across settings in Africa, and examine how other educational interventions affect teacher effectiveness.

References

Altinyelken, Hulya Kosar (2010). "Curriculum Change in Uganda: Teacher Perspectives on the New Thematic Curriculum". *International Journal of Educational Development* 30.2, pp. 151–161. DOI: 10.1016/j.ijedudev.2009.03.004.

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.

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.

Brown, Byron W. and Daniel H. Saks (1981). "The Microeconomics of Schooling". *Review of Research in Education* 9. Publisher: [Sage Publications, Inc., American Educational Research Association], pp. 217–254. ISSN: 0091-732X. DOI: 10.2307/1167186.

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.

Condie, Scott, Lars Lefgren, and David Sims (2014). "Teacher heterogeneity, value-added and education policy". *Economics of Education Review* 40, pp. 76–92. ISSN: 02727757. DOI: 10.1016/j.econedurev.2013.11.009.

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.

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.

Gilligan, Daniel O, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M Lucas, and Derek Neal (2018). *Educator incentives and educational triage in rural primary schools*. National Bureau of Economic Research.

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.

Gray-Lobe, Guthrie, Anthony Keats, Michael Kremer, Isaac Mbiti, and Owen W. Ozier (2022). *Can Education be Standardized? Evidence from Kenya*. SSRN Scholarly Paper 4129184. Rochester, NY.

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*.

Kerwin, Jason T. and Rebecca L. Thornton (2021). "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures". *The Review of Economics and Statistics* 103.2, pp. 251–264. DOI: 10.1162/rest_a_00911.

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, Julian R Betts, et al. (2007). *Re-examining the role of teacher quality in the educational production function*. National Center on Performance Incentives, Vanderbilt, Peabody College.

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.

Loeb, Susanna, Demetra Kalogrides, and Tara Béteille (2012). "Effective schools: Teacher hiring, assignment, development, and retention". *Education Finance and Policy* 7.3. Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . ., pp. 269–304.

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.

— (2017). "Measuring the Impacts of Teachers: Comment". *American Economic Review* 107.6, pp. 1656–1684. ISSN: 0002-8282. DOI: 10.1257/aer.20141440.

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.

Ssentanda, Medadi Erisa, Kate Huddlestone, and Frenette Southwood (2016). "The Politics of Mother Tongue Education: The Case of Uganda". *Per Linguam* 32.3, pp. 60–78. DOI: 10.5785/32-3-689.

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.

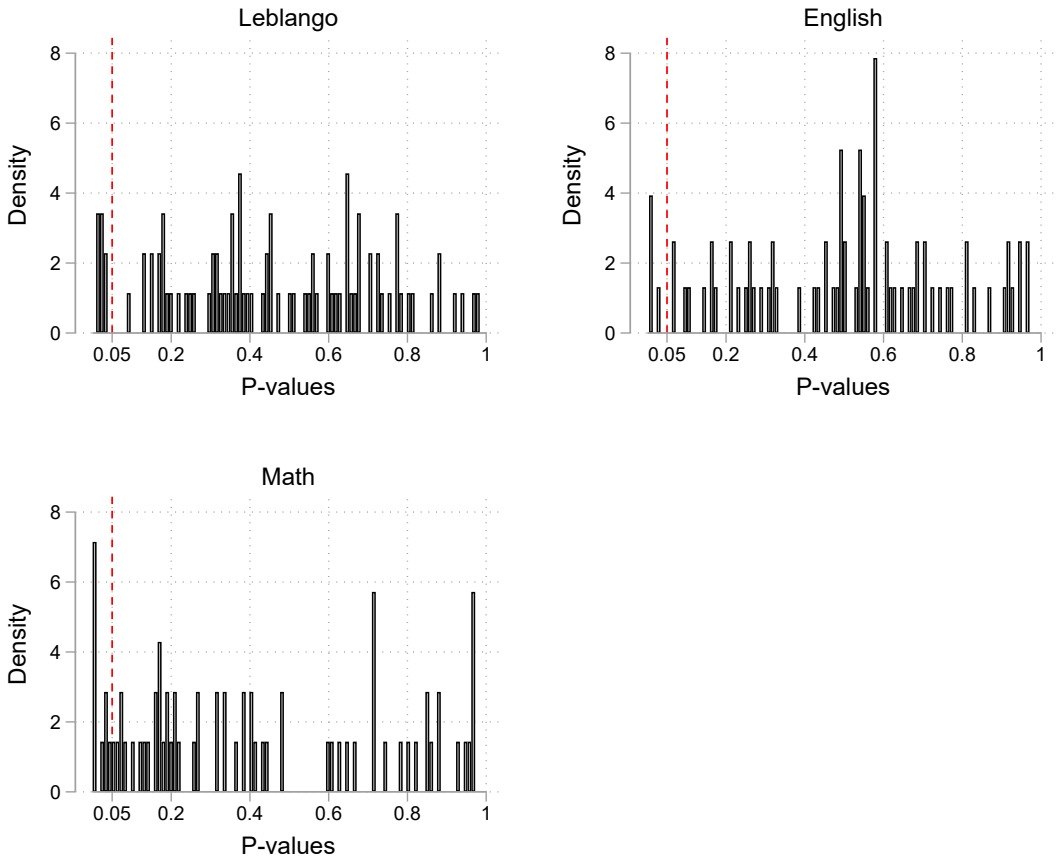
Uwezo (2016). *Are Our Children Learning (2016)? Uwezo Uganda Sixth Learning Assessment Report*. Kampala: Twaweza East Africa.

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

A Online Appendix

A.1 Appendix Figures

Appendix Figure A1
Horvarth (2015) Test



Notes: The panels in this figure graph the p-values of testing differences in average baseline test scores between classrooms within grades and schools within each of the 42 control schools. We use data from years with random assignment only (2013, 2016 and 2017). The red vertical line mark a p-value of 0.05..

A.2 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
Math tests (all grades)	Endline	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
NULP Sample Across Study Arms

	All	Control	Reduced Cost	Full Cost
Panel A: NULP Evaluation Sample				
Schools	128	42	44	42
Teachers	1,480	497	521	462
Classrooms	2,314	763	798	753
Students with at least one test	28,533	9,163	9,983	9,387
Student-year obs with at least one test	60,681	19,238	20,993	20,450
Teachers/Grade	2.34	2.38	2.29	2.36
Panel B: Students with Consecutive Tests				
Student-year obs with endline EGRA test	58,774	18,639	20,416	19,719
Student-year obs with consecutive Leblango tests	49,054	15,427	17,037	16,590
Student-year obs with endline English test	37,077	11,718	12,816	12,543
Student-year obs with consecutive English tests	27,300	8,493	9,408	9,399
Student-year obs with endline Math test	57,005	18,024	19,650	19,331
Student-year obs with consecutive Math tests	48,731	15,264	16,669	16,798
Panel C: Matching Students to Teachers				
Student-year obs matched to a teacher	58,154	18,521	20,041	19,592
Student-year obs in a class size > 5 (Leblango)	55,702	17,571	19,209	18,922
Student-year obs in a class size > 5 (English)	34,722	10,865	11,928	11,929
Student-year obs in a class size > 5 (Math)	54,946	17,422	18,820	18,704

Notes: The 128 schools were sampled in two phases: 38 in 2013 and additional 90 in 2014.

Appendix Table A3

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 A4
Descriptive Statistics across Treatment Arms

	Control	**Reduced-Cost**	**Full-Cost**	*p*-value from F-test between study arms
	(1)	(2)	(3)	(4)
Panel A: Students				
Female (%)	0.497	0.508	0.496	0.72
Age	8.944	8.992	8.999	0.76
BL Leblango (P1)	–0.037	–0.045	–0.004	0.23
BL Math (P1)	–0.039	–0.051	–0.062	0.68
Panel B: Teachers				
Women (%)	0.464	0.448	0.397	0.07
Age	39.593	40.133	39.530	0.27
Yrs Experience	14.197	14.131	14.306	0.71
Yrs of education	15.623	15.542	15.504	0.95
Education Level				
UACE or less	0.006	0.037	0.021	0.14
Certificate	0.674	0.673	0.732	0.64
Diploma	0.308	0.281	0.241	0.75
Degree	0.012	0.009	0.006	0.92
Teachers with characteristics data	319	334	318	

Notes: Column 4 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	0.014***	0.003	0.001
	(0.005)	(0.004)	(0.004)	(0.005)
Female × Reduced-cost				0.013*
				(0.006)
Female × Full-cost				0.002
				(0.007)
Age	-0.014***	-0.011***	-0.010***	-0.014***
	(0.003)	(0.002)	(0.002)	(0.003)
Age × Reduced-cost				0.002
				(0.004)
Age × Full-cost				0.004
				(0.004)
Grade Level (expected)	0.113***	0.098***	0.101***	0.110***
	(0.007)	(0.006)	(0.007)	(0.006)
Grade Level × CCT				-0.012*
				(0.006)
Grade Level × MT				-0.005
				(0.007)
Reduced-cost program				-0.037
				(0.029)
Full-cost program				-0.062**
				(0.029)
Observations	23,663	25,675	24,686	74,024
Adjusted R-squared	0.051	0.048	0.039	0.046

Notes: Attrition 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 Student Attrition and Teacher Characteristics

Teacher characteristics	**Control** (1)	**Reduced-cost** (2)	**Full-cost** (3)	**All** (4)
Female (1=Yes)	-0.004**	-0.001	-0.003	-0.006***
	(0.002)	(0.003)	(0.005)	(0.002)
Female × Reduced-cost				0.006*
				(0.004)
Female × Full-cost				0.006
				(0.006)
Experience (years)	-0.000	0.000	0.000	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
Experience × Reduced-cost				0.001***
				(0.000)
Experience × Full-cost				0.001**
				(0.000)
> Certificate (1=Yes)	0.005	-0.008	-0.001	0.008**
	(0.003)	(0.005)	(0.004)	(0.003)
> Certificate × Reduced-cost				-0.017**
				(0.007)
> Certificate × Full-cost				-0.008
				(0.005)
Reduced-cost program				-0.004
				(0.004)
Full-cost program				-0.010***
				(0.004)
Observations	16,537	17,647	17,777	51,961
Adjusted R-squared	0.127	0.143	0.086	0.110

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

Appendix Table A7
Correlation between Teacher Attrition and Teacher Characteristics

Teacher characteristics	**Control** (1)	**Reduced-cost** (2)	**Full-cost** (3)	**All** (4)
Female (1=Yes)	-0.093*	0.078	-0.045	-0.068
	(0.054)	(0.052)	(0.047)	(0.046)
Female × Reduced-cost				0.157**
				(0.064)
Female × Full-cost				0.020
				(0.061)
Experience (years)	0.005	0.002	0.003	0.004
	(0.004)	(0.003)	(0.003)	(0.004)
Experience × Reduced-cost				-0.002
				(0.005)
Experience × Full-cost				-0.002
				(0.004)
> Certificate (1=Yes)	0.005	-0.053	-0.006	0.060
	(0.076)	(0.066)	(0.063)	(0.057)
> Certificate (1=Yes) × Reduced-cost				-0.134*
				(0.078)
> Certificate (1=Yes) × Full-cost				-0.088
				(0.078)
Reduced-cost program				-0.008
				(0.075)
Full-cost program				0.011
				(0.074)
Observations	319	334	318	971
Adjusted R-squared	0.012	0.013	0.030	0.006

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

Appendix Table A8
Correlation between being an Incoming Teacher and Teacher Characteristics

Teacher characteristics	**Control** (1)	**Reduced-cost** (2)	**Full-cost** (3)	**All** (4)
Female (1=Yes)	-0.091**	-0.114**	-0.198***	-0.092***
	(0.040)	(0.049)	(0.051)	(0.031)
Female × Reduced-cost				-0.013
				(0.049)
Female × Full-cost				-0.054
				(0.052)
Experience (years)	-0.006*	-0.015***	-0.008**	-0.006***
	(0.003)	(0.003)	(0.003)	(0.002)
Experience × Reduced-cost				-0.007**
				(0.004)
Experience × Full-cost				-0.001
				(0.004)
> Certificate (1=Yes)	0.098	0.090	0.097	0.088**
	(0.061)	(0.064)	(0.074)	(0.042)
> Certificate (1=Yes) × Reduced-cost				-0.003
				(0.059)
> Certificate (1=Yes) × Full-cost				0.007
				(0.068)
Reduced-cost program				0.093
				(0.058)
Full-cost program				-0.032
				(0.056)
Observations	319	334	318	971
Adjusted R-squared	-0.094	-0.041	-0.053	0.040

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

Appendix Table A9

Tests Used to Estimate Value-Added

		2013	2014	2015	2016	2017
Panel A: Leblango Reading and Math						
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 A10
Correlation between Student and Teacher Characteristics

	(1) Female	(2) Age	(3) BL Leblango	(4) BL English	(5) BL Math
Teacher characteristics					
Age	0.001	0.004	-0.004	-0.009	-0.005
	(0.001)	(0.003)	(0.004)	(0.008)	(0.004)
Female	0.012	-0.023	-0.012	-0.007	0.005
	(0.009)	(0.024)	(0.026)	(0.039)	(0.026)
Experience	-0.000	-0.002	0.004	0.008	0.008*
	(0.001)	(0.003)	(0.004)	(0.008)	(0.004)
> Certificate	0.001	-0.078*	0.041	0.106*	-0.037
	(0.013)	(0.040)	(0.037)	(0.058)	(0.036)
Observations	16,185	16,071	16,191	16,191	16,191
Adjusted R-squared	0.003	0.536	0.039	0.206	0.055

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

Appendix Table A11

Classroom and Teacher Value-Added: Control Schools

	Leblango		English		Math	
	Classroom	**Teacher**	**Classroom**	**Teacher**	**Classroom**	**Teacher**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Including School Effects						
SD	0.41	0.29	0.54	0.45	0.55	0.44
	[0.29,0.53]	[0.19,0.39]	[0.34,0.74]	[0.27,0.63]	[0.46,0.64]	[0.37,0.51]
Corrected SD	0.36	0.27	0.52	0.43	0.51	0.42
	[0.22,0.50]	[0.17,0.38]	[0.31,0.73]	[0.25,0.62]	[0.42,0.60]	[0.35,0.49]
Observations	17,571	11,673	10,865	6,333	17,422	11,568
Classrooms/Teachers	571/361	362/152	392/285	219/112	571/361	362/152
Schools	42	42	42	42	42	42
Panel B: School Effects Purged						
SD	0.39	0.26	0.35	0.24	0.45	0.31
	[0.26,0.52]	[0.16,0.36]	[0.21,0.49]	[0.08,0.40]	[0.34,0.56]	[0.21,0.41]
Corrected SD	0.33	0.24	0.31	0.22	0.41	0.30
	[0.18,0.48]	[0.14,0.34]	[0.17,0.45]	[0.05,0.39]	[0.29,0.52]	[0.20,0.40]
Observations	17,571	11,673	10,865	6,333	17,422	11,568
Classrooms/Teachers	571/361	362/152	392/285	219/112	571/361	362/152
Schools	42	42	42	42	42	42
Panel C: School-by-grade Effects Purged						
SD	0.22	0.11	0.23	0.12	0.35	0.19
	[0.17,0.27]	[0.10,0.12]	[0.19,0.27]	[0.08,0.16]	[0.27,0.43]	[0.16,0.22]
Corrected SD	0.11	0.09	0.17	0.11	0.30	0.18
	[0.08,0.14]	[0.08,0.10]	[0.13,0.21]	[0.07,0.15]	[0.23,0.37]	[0.15,0.21]
Observations	14,202	8,784	8,757	4,777	14,073	8,698
Classrooms/Teachers	491/322	293/124	338/253	178/93	491/322	293/124
Schools	42	39	42	37	42	39
Panel D: Random Assignment Years - School-by-grade Effects Purged						
SD	0.21	0.11	0.21	0.13	0.29	0.19
	[0.19,0.23]	[0.10,0.12]	[0.19,0.23]	[0.11,0.15]	[0.27,0.31]	[0.17,0.21]
Corrected SD	0.12	0.09	0.15	0.12	0.24	0.18
	[0.10,0.14]	[0.09,0.09]	[0.13,0.17]	[0.11,0.13]	[0.22,0.26]	[0.16,0.20]
Observations	5,963	2,315	4,647	1,672	5,915	2,289
Classrooms/Teachers	244/199	90/45	190/158	65/33	244/199	90/45
Schools	36	22	36	22	36	22

Notes: 95% confidence intervals for the SD of the classroom/teacher effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A12

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

	Leblango	Leblango	English	English	Math	Math
	Classroom Effects	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Same Sample of Teachers between Classroom and Teacher Samples						
Corrected SD	0.14	0.09	0.21	0.11	0.35	0.18
	[0.09,0.19]	[0.06,0.12]	[0.13,0.29]	[0.08,0.14]	[0.25,0.45]	[0.14,0.22]
Observations	8,784	8,784	4,777	4,777	8,698	8,698
Classrooms/Teachers	293/124	293/124	178/93	178/93	293/124	293/124
Schools	39	39	37	37	39	39
Panel B: Same Sample of Teachers between School and Grade Sample						
Corrected SD	0.32	0.25	0.30	0.22	0.42	0.31
	[0.19,0.45]	[0.15,0.35]	[0.21,0.39]	[0.13,0.31]	[0.30,0.54]	[0.22,0.40]
Observations	14,202	8,784	8,757	4,777	14,073	8,698
Classrooms/Teachers	491/322	293/124	338/253	178/93	491/322	293/124
Schools	42	39	42	37	42	39

Notes: 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-by-grade effects by subtracting off the school-by-grade mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A13

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

	Leblango	Leblango	English	English	Math	Math
	Classroom Effects	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Minimum of 10 Students						
Corrected SD	0.12	0.10	0.17	0.10	0.29	0.21
	[0.07,0.17]	[0.08,0.12]	[0.11,0.23]	[0.06,0.14]	[0.21,0.37]	[0.15,0.27]
Observations	13,663	8,499	8,367	4,604	13,538	8,415
Classrooms/Teachers	422/287	259/124	285/217	155/87	422/287	259/124
Schools	42	39	42	36	42	39
Panel B: Minimum of 15 Students						
Corrected SD	0.09	0.07	0.16	0.11	0.27	0.22
	[0.05,0.13]	[0.05,0.09]	[0.11,0.21]	[0.08,0.14]	[0.21,0.33]	[0.16,0.28]
Observations	12,866	8,018	7,807	4,308	12,751	7,939
Classrooms/Teachers	359/253	221/115	239/186	131/78	359/253	221/115
Schools	41	38	41	36	41	38

Notes: 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-by-grade effects by subtracting off the school-by-grade mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A14

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

	Leblango	Leblango	English	English	Math	Math
	Classroom Effects	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Omitting student-year observations with missing characteristics						
Corrected SD	0.13	0.12	0.19	0.12	0.29	0.19
	[0.08,0.19]	[0.09,0.14]	[0.13,0.25]	[0.09,0.16]	[0.24,0.34]	[0.16,0.22]
Observations	11,516	7,226	6,071	3,219	11,437	7,172
Classrooms/Teachers	473/314	283/124	319/244	168/93	473/314	283/124
Schools	42	39	42	37	42	39
Panel B: Omitting grade-one student-year observations						
Corrected SD	0.12	0.08	0.17	0.11	0.29	0.16
	[0.08,0.16]	[0.06,0.10]	[0.12,0.22]	[0.07,0.15]	[0.19,0.39]	[0.12,0.20]
Observations	8,757	4,777	8,757	4,777	8,646	4,703
Classrooms/Teachers	338/253	178/93	338/253	178/93	338/253	178/93
Schools	42	37	42	37	42	37

Notes: 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-by-grade effects by subtracting off the school-by-grade mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A15
Robustness Estimates of Teacher Value-Added: Using Alternative Outcomes, Control Schools

	Leblango	Leblango	English	English	Math	Math
	Classroom Effects	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**	**Classroom Effects**	**Teacher Effects**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Gain Score Model						
Corrected SD	0.11	0.09	0.16	0.11	0.35	0.20
	[0.05,0.17]	[0.07,0.11]	[0.08,0.24]	[0.08,0.14]	[0.25,0.45]	[0.17,0.23]
Observations	14,202	8,784	8,757	4,777	14,073	8,698
Classrooms/Teachers	491/322	293/124	338/253	178/93	491/322	293/124
Schools	42	39	42	37	42	39
Panel B: Simple Index						
Corrected SD	0.11	0.09	0.18	0.13	0.32	0.19
	[0.07,0.15]	[0.07,0.11]	[0.13,0.23]	[0.06,0.20]	[0.26,0.38]	[0.16,0.22]
Observations	14,202	8,784	8,757	4,777	14,073	8,698
Classrooms/Teachers	491/322	293/124	338/253	178/93	491/322	293/124
Schools	42	39	42	37	42	39

Notes: 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-by-grade effects by subtracting off the school-by-grade mean. Control schools (N=42) did not receive the NULP intervention.

Appendix Table A16
Average Treatment Effects of the NULP

	(1) Leblango	(2) Leblango
Reduced-cost	0.691***	0.220***
	(0.096)	(0.060)
Full-cost	1.209***	0.540***
	(0.106)	(0.071)
Observations	6,118	45,436
R-squared	0.130	0.095
Sample	Cohort 2 after 3 years of exposure	All students across all years

Notes: All regressions include controls for age, gender and dummy variables indicating if these are missing, as well as stratification cell fixed effects. Standard errors clustered at the school-level in parentheses.

Appendix Table A17

Robustness of 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
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD	0.03	0.28	0.31	0.09	0.18	0.18
	[–0.03,0.09]	[0.21,0.35]	[0.25,0.37]	[0.06,0.12]	[0.11,0.25]	[0.12,0.24]
Observations	11,494	13,244	12,478	7,738	9,720	9,359
Classrooms/Teachers	376/250	425/266	389/239	250/124	300/141	288/138
Schools	41	42	41	39	41	37

Notes: All estimates are calculated using data between 2013 and 2016. All estimates are purged of school-by-grade effects by subtracting off the school-by-grade 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 A18

Robustness of 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
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD	0.13	0.24	0.31	0.10	0.19	0.15
	[0.07,0.19]	[0.17,0.31]	[0.25,0.37]	[0.07,0.13]	[0.09,0.29]	[0.12,0.18]
Observations	9,919	11,710	11,690	7,141	9,304	9,587
Classrooms/Teachers	314/189	371/204	357/187	227/102	293/126	295/125
Schools	41	42	41	37	41	37

Notes: All estimates are calculated using teachers teaching the treated cohorts; P1 (2013 and 2014), P2 (2015), and P3 (2016) as well as the classes they taught after. All estimates are purged of school-by-grade effects by subtracting off the school-by-grade 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 A19

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

	Classroom Effects		
	Control	Reduced-Cost	Full-Cost
	(1)	(2)	(3)
Corrected SD	0.07	0.30	0.32
	[0.00,0.14]	[0.21,0.39]	[0.26,0.38]
Observations	6,140	6,643	6,556
Classrooms/Teachers	192/167	208/181	197/164
Schools	41	42	41

Notes: All estimates are calculated using only teachers teaching the treated cohorts; P1 (2013 and 2014), P2 (2015), and P3 (2016). All estimates are purged of school-by-grade effects by subtracting off the school-by-grade 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 A20

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

	Classroom Effects			Teacher Effects		
	Control	Reduced-Costs	Full-Costs	Control	Reduced-Costs	Full-Costs
	(1)	(2)	(3)	(4)	(5)	(6)
Corrected SD	0.15	0.18	0.34	0.10	0.10	0.14
	[0.08,0.22]	[0.10,0.26]	[0.25,0.43]	[0.07,0.13]	[0.08,0.13]	[0.12,0.16]
Observations	4,068	3,911	5,221	3,940	3,865	5,003
Classrooms/Teachers	119/49	127/47	156/60	114/44	124/44	150/54
Schools	28	25	31	25	24	27

Notes: All estimates are calculated using only teachers treated by the NULP in multiple years. All estimates are purged of school-by-grade effects by subtracting off the school-by-grade mean. 95% confidence intervals for the SD of the classroom effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications.