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*

November 28, 2024

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, but do find important sorting of teachers to grades. 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 and NBER (econjeff@ssc.wisc.edu) Thornton: Department of Economics, Hankamer School of Business, Baylor University (Rebecca_Thornton@baylor.edu). We thank Laura Schechter, Chao Fu, and seminar audiences at the University of Minnesota, Michigan State University, 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.[1] Hence, if teachers vary substantially in their contribution 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, often interpreted as the scope for policies to improve student learning with interventions targeting teachers (e.g. Rivkin, Hanushek, and Kain 2005; Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014). Implicit in the 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 to improve teacher quality. Existing research focuses almost entirely 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 the gap in the literature by first estimating the variance of teacher effectiveness and then measuring the causal effect of a teacher-focused intervention on the distribution of teacher effectiveness.

We first present 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. Next, guided by our model of learning and teacher effectiveness, we use longitudinal data from schools, teachers, and students in Northern Uganda to obtain the first teacher value-added estimates in sub-Saharan Africa. Lastly, we 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 empirical results involve estimating lower bounds on the variation of classroom and teacher effectiveness (measured as value-added).[2] These results use data

[1] See for example, 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), and McEwan (2015).

[2] We use "effectiveness" and "value-added" interchangeably throughout the paper. A related literature examines the value-added of schools rather than teachers or classrooms. 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 post-dating our study, Oketch, Rolleston, and Rossiter (2021) estimate classroom value-added in Ethiopia, but do not isolate the stable contributions of specific teachers.

from 2013 to 2017 from 42 control schools that did not receive the NULP. We distinguish between "classroom effects", which are the causal effect of being in a specific classroom in a given year, and "teacher effects", the stable component of classroom effects attributable to a given teacher. We estimate classroom value-added using student test scores where we observe at least two teachers per school; we estimate teacher value-added among teachers who teach across multiple years in our data. We present several sets of estimates: within-school estimates that do not account for sorting of teachers to school or grade-levels, and within-school and within school-grade estimates that do.

Our within-school estimates indicate 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. However, we show that in our setting there is systematic sorting of teachers to grades. Our preferred estimates, therefore, focus on variation within grade and school and are 0.09 SDs for local-language reading, 0.11 for English, and 0.18 for Math. To address the potential bias arising from non-random sorting of students to classrooms, we utilize the fact that three of the five years of data collection involved randomly assigning 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.[3] 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 other settings.[4] Our estimates of teacher and classroom value-added are positively correlated across subjects, suggesting that teachers are broadly effective across subjects, consistent with prior literature (Koedel, Betts, et al. 2007; Loeb, Kalogrides, and Béteille 2012; Goldhaber, Cowan, and Walch 2013; Condie, Lefgren, and Sims 2014).

The second set of empirical results in the paper involves estimating the causal effect of

[3] On the one hand, a difficult teaching environment may amplify the importance of teachers. On the other hand, difficult conditions may render even the strongest teachers ineffective. Data from sub-Saharan Africa paint a bleak picture of the school teaching environment. For example, Bold et al. (2017) find that "essentially no public primary schools in... [Kenya, Mozambique, Nigeria, Senegal, Tanzania, Togo, and Uganda] offer adequate quality education".

[4] Hanushek and Rivkin (2012) provide an average estimate of the variance of teacher value-added for reading of 0.13 SDs based on nine studies in United States schools, while Sass et al. (2012) find greater variation in teacher value-added in high poverty schools than lower poverty ones in North Carolina and Florida. Chetty, Friedman, and Rockoff (2014) estimate a teacher value-added of 0.10 SDs. Araujo et al. 2016 estimate a SD of 0.09 among kindergarten teachers in Ecuador, and Bau and Das (2020) estimate a SD of 0.06 in Pakistan. 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.

an educational intervention on the distribution of teacher value-added. The NULP provided intensive training and support to teachers in grades one to three in literacy instruction, with a focus on local-language reading. We utilize the random assignment of our sample schools to three treatment arms: the 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 caused students in full-cost program schools to score 1.21 SDs higher and students in reduced-cost program schools to score 0.69 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 the focus of the NULP intervention was local-language literacy. We estimate value-added separately in each treatment arm and find that both versions of the intervention increase the spread of the distribution of classroom and teacher effectiveness. The program increased the distribution of teacher effects 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 presented in Section 2, shows how testing for rank preservation in teacher quality across intervention treatment arms can 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. Rejecting rank preservation implies that at least some teachers become relatively less effective. It allows, but by no means implies, a negative correlation between value-added in the treated and baseline states. We formally test for rank preservation following Bitler, Gelbach, and Hoynes (2005) and Djebbari and Smith (2008), and fail to reject the null hypothesis of rank preservation. This suggests that the NULP likely had the largest effects amongst the most-effective teachers.

Lastly, we examine how teacher characteristics correlate with our estimated measures of value-added, and the gains in value-added resulting from the NULP. Prior research (in both developed and developing country contexts) finds that, among commonly observed teacher characteristics, being a new teacher seems to meaningfully predict teacher value-added (e.g. 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 evidence 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 our conceptual framework for thinking about how the distribution of teacher value-added varies with educational inputs and how training teachers might affect that distribution. Section 3 describes the setting, details about the NULP intervention and evaluation, and descriptive statistics related to the sorting of teachers to schools, grades and classrooms. In Section 4, we provide a description of our analytical samples including a discussion of balance and attrition. 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 as well as present tests for rank preservation and correlations with teacher characteristics. Throughout the two results sections we present robustness and sensitivity analysis. 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 Achievement

Consider a model of academic achievement for student i, in school s, in grade g, in classroom c, with teacher j, learning subject k, which we denote by Y^k_{isgcj}, where $k \in \{math, reading\}$.[5,6] To reduce notation at the margin, in this section we imagine a single cohort of students and so omit calendar time.[7]

In our model, as in life, achievement in grade g depends on the entire sequence of inputs at home and in school. Let H_{ig} denote home inputs to student i in grade g, and let $H_i(g)$ denote the vector of such inputs in all periods up to and including grade g. We conceive of home inputs as encompassing the time and goods provided by parents and others outside the school, ranging from books to healthcare to instruction in football. Achievement also depends on the student's fixed subject-specific genetic endowment (θ^k_i).

Given our focus on teachers, we model educational inputs in greater detail than home inputs. School-level inputs comprise those experienced by *all* students within a school such

[5] We use grade rather than age because not all students in our context start school at the same age.

[6] We add the complication that students learn to read both Leblango and English later on.

[7] Buhl-Wiggers et al. (2024) show that the students we study experience substantial grade repetition, so time relative to schooling start often differs from grade level. We ignore this complication for now.

as the head teacher and their staff, the physical environment of the school and its grounds and so on. We denote school-level inputs in grade g relevant to the production of achievement in subject k by S_{sg}^{ik}, and the corresponding vector of such inputs up to and including grade g experienced by student i by $S_{is}(g)$. We call the effect of classroom c in school s on achievement in subject k in grade g the "classroom effect" and write it C_{csg}^{k}. Let $C_{ics}^{k}(g)$ denote the sequence of classroom effects up to and including grade g experienced by student i. Following the literature, we assume a common classroom effect, implying no i subscript, but we do put an i subscript on the sequence, which will vary among students.

With this notation in hand, we can write the production function for student achievement in terms of the classroom effects and other inputs as:

$$Y_{isg}^{k} = g[H_i(g), S_{is}(g), C_{icst}^{k}(g), \theta_i^{k}]. \tag{1}$$

Three features of this formulation merit special attention. First, while the inputs vary among students, the production function does not; more prosaically, the $g[.]$ function lacks an i subscript. This represents a strong, substantive restriction. Second, it emphasizes the cumulative nature of achievement production. An ideal analysis would have data on all inputs in every period up to the time of outcome measurement, which would allow direct estimation of a version of Equation (1). Finally, this formulation assumes that the classroom effects in one subject do not affect outcomes in the other. We relax this assumption in a "reduced form" way in our empirical setup in Section 5.

We, like nearly all other researchers, lack the cumulative data necessary to directly estimate a version of Equation (1). Instead, we thoughtfully follow the literature by letting prior student achievement (Y_{isjt-1}^{k}) proxy for the unobserved input histories as well as the student's genetic endowment. Todd and Wolpin (2003) provide a formal justification for this strategy. Adopting this strategy, we can rewrite the production function in terms of current inputs and lagged achievement as follows:

$$Y_{isg}^{k} = l\{Y_{isg-1}^{k}[H_i(g-1), S_{is}(g-1), C_{ics}^{k}(g-1), \theta_i^{k}], H_i, S_{is}, C_{icsg}^{k}\}. \tag{2}$$

In Section 5, we transform Equation (2) into an estimating equation by boldly assuming additive separability in all of the inputs, by including some additional covariates to add flexibility to the empirical production function, and by adding an additively separable "error" term that captures both measurement error in the test scores that we use as measures of achievement, as well as errors of functional form and the remaining unobserved determinants of student achievement.

2.2 On the Origin of Classroom Effects

Talking about the NULP intervention within our framework necessitates further refinements. In particular, we need to say more about where our classroom effects come from. To that end, let the classroom effect depend on the materials M^k_{csg} allocated to it, the input of the teacher leading it J^k_{cst}, and an aggregate of other factors, R^k_{csg}, that includes the students themselves as well as various unobserved factors specific to the classroom as experienced by a given cohort of students.[8] In notation, $C^k_{csg} = h(M^k_{csg}, J^k_{csg}, R^k_{csg})$. We assume positive first derivatives and negative second derivatives, so that each input increases student achievement albeit at a decreasing rate, holding fixed the other inputs.[9] On classroom materials, see e.g. Brown and Saks (1981) for empirical evidence supporting this functional form; for teachers and "other" inputs ours assumptions have a more definitional flavor.

Drilling down a bit further, we allow the inputs from each teacher j to depend on their effort, given by E^k_{sjg}, and on their skills, given by L^k_{sjg}. Put differently, $J^k_{csg} = J(E^k_{sjg}, L^k_{sjg})$. We again assume positive first derivatives and negative second derivatives, but take no stand on the cross-partial derivative. Put differently, we leave the complementarity or substitutability of teacher skill and effort for future research to sort out. The teacher inputs have corresponding vector sequences $E^k_{sj}(g)$ and $L^k_{sj}(g)$.

We allow teacher effort to vary over time, reflecting things like health shocks and distractions at home, and we allow teacher skills to change over time in response to, among other things, the intervention we study. Teacher inputs also potentially vary by subject. Teachers may have more skills teaching one subject than another, perhaps due to differences in their educational paths, perhaps due to post-schooling investments in professional development, or perhaps due to natural differences in aptitude. Teachers will put different amounts of effort into each subject depending on their perceptions of student needs and abilities, their perceptions of their own skills, and which subject they enjoy teaching more. We imagine a teacher effort budget hiding in the background, leading teachers who put a lot of effort into one subject to (in general) put less into the other.

Substituting for C^k_{csg} and then for J^k_{csg} in Equation (1) yields

$$Y^k_{isjg} = g[H_i(g), S_{is}(g), M^k_{sj}(g)J(E^k_{sj}(g), L^k_{sj}(g)), \theta^k_i] \tag{3}$$

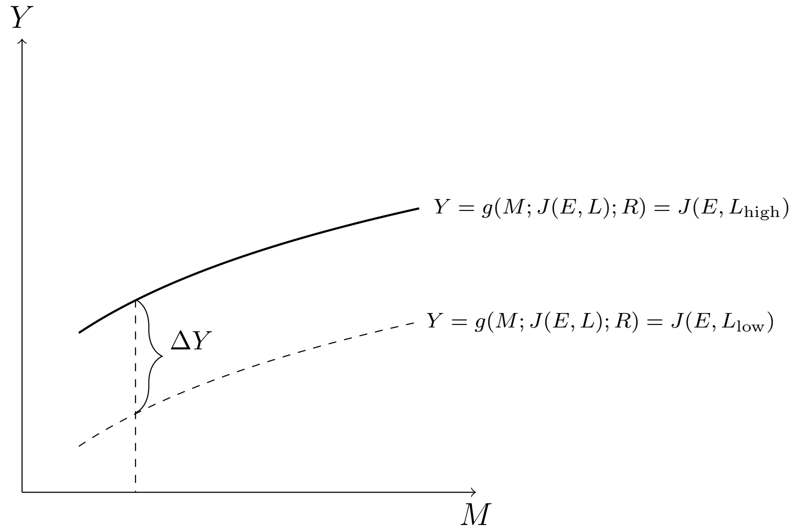
[8] For example, one semester one of us taught undergraduate econometrics at just the time when university grounds staff noisily mowed the lawn outside a specific classroom each week.

[9] 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.

We describe the NULP intervention in loving detail in Section 3.2. For the moment, it suffices to note that it includes both classroom materials and teacher training in how to use the materials in their teaching. As its name suggests, the intervention aims at reading, rather than math. As such, it directly increases $M_{sj}^k(g)$ and $L_{sj}^k(g))$ for $k = reading$. In contrast, its effect on teacher effort $E_{sj}^k(g)$ depends on the cross-partial between effort and skills as well as on the implicit "income effect" that results from other inputs reducing the effort price of any particular level of student achievement. Similar ambiguity surrounds the NULP intervention's effect on math, with the added complications that some of the materials in offers (e.g., the clocks) should improve learning in all subjects and that improvements in students reading may make learning math easier for them.

Figure 1 illustrates the direct effects of classroom materials and teacher skills on student outcomes. The horizontal axis varies the amount of classroom materials. The vertical axis indicates student reading skills. The upper (solid) line shows how student skills increase with classroom materials (albeit at a decreasing rate), holding teacher effort E constant and fixing teacher skills at a high level (L_{high}), where we simplify our earlier notation for the figure in an obvious way. Similarly, the lower (dashed) curve shows the relationship between classroom materials and student skills fixing teacher skills at a low level (L_{low}). Holding materials fixed at any point on the horizontal access and jumping from the low curve to the higher curve ilustrates the increment to student skills (ΔY) associated with an increase in teacher skill from L_{low} to L_{high}. In the figure, the NULP intervention leads to both a higher curve, by improving teacher skills, and generates a move to the right along the curve, by improving classroom materials.

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


Y
$Y = g(M; J(E, L); R) = J(E, L_{\text{high}})$
ΔY
$Y = g(M; J(E, L); R) = J(E, L_{\text{low}})$
M

2.3 Teacher Value-Added

Heretofore, we have considered classroom effects $C^k_{csg} = h(M^k_{csg}, J^k_{csg}, R^k_{csg})$ that capture the increment to students' outcomes associated with a particular classroom in a particular school in a particular grade. These classroom effects vary both cross-sectionally and temporally due to differences in teacher skills, teacher effort, and classroom materials. To learn more about teachers, we now consider how to combine the classroom effects associated with specific teachers to produce a teacher effect.

To simplify the discussion, assume additive separability, so that

$$C^k_{csg} = h_M(M^k_{sj}) + h_J(J^k_{csg}) + h_R(R^k_{csg}) \tag{4}$$

where the sub-functions retain their positive first and negative second derivatives but the cross-partials now all equal zero by assumption. We (and the literature) define the teacher effect as the persistent component of the teacher input into the classroom effect. In notation, the teacher effect becomes

$$T = E_g(h_J(J^k_{csg})) \tag{5}$$

where the expectation is across grades within teacher. Our interest in teacher effects depends in a sense on the stability of the skill and effort input decisions that teachers make. Should effort matter a lot, and should teachers vary their effort a lot across years, then the persistent component of $h_J(J^k_{csg})$. Teacher skills likely move more slowly, but the same point applies. We return to these issues in our discussion of our empirical estimates of the variance of teacher value-added and when we examine the extent to which various teacher characteristics (such as experience, which proxies for skill) correlate with teacher value-added.[10]

As described with greater specificity in Section 5, we follow the literature and estimate the teacher effect as a (conditional) average of classroom effects associated with a given teacher. Some reflection on Equation (4) makes clear that, to the extent that particular teachers have persistent differences in materials, whether because they purchase them out of their own funds or because they excel at extracting them from the head teacher, the teacher effect will incorporate the effects of those materials.[11] For this reason, the treatment effects we estimate for teachers treated by the NULP intervention in Section 6 will incorporate any effects of the materials it provides, along with any ongoing effects on teacher effort and skills.

Equation (5) defines a teacher effect for any subject k. Below, we estimate teacher effects in both math and reading and examine their relationship. This exercise reveals something about the relative importance of absolute advantage and comparative advantage. Under absolute advantage, teachers are broadly effective (or broadly ineffective) across multiple subjects, implying a positive correlation across subject-specific value added. In contrast, under comparative advantage, teachers who excel in one subject lag behind in others, implying a negative correlation.[12]

Some remarks: First, in practice, we need at least two classroom observations to estimate a teacher effect. We will drop teachers we only see once in our empirical work. Second, the literature refers to teacher effects that capture the persistent component of teacher inputs into classroom effects as “teacher value added.” We will use “teacher effect”, “teacher effectiveness” and “teacher value-added” as synonyms for the remainder of the paper. Third, readers familiar with the literature on teacher value-added will anticipate that we address a number of applied econometric points later on in Section 5.

[10] The developed country literature finds that teacher value-added tends to evolve slowly, other than a quick rise in the first few years of teaching.

[11] The same point applies to persistent differences in “other factors”.

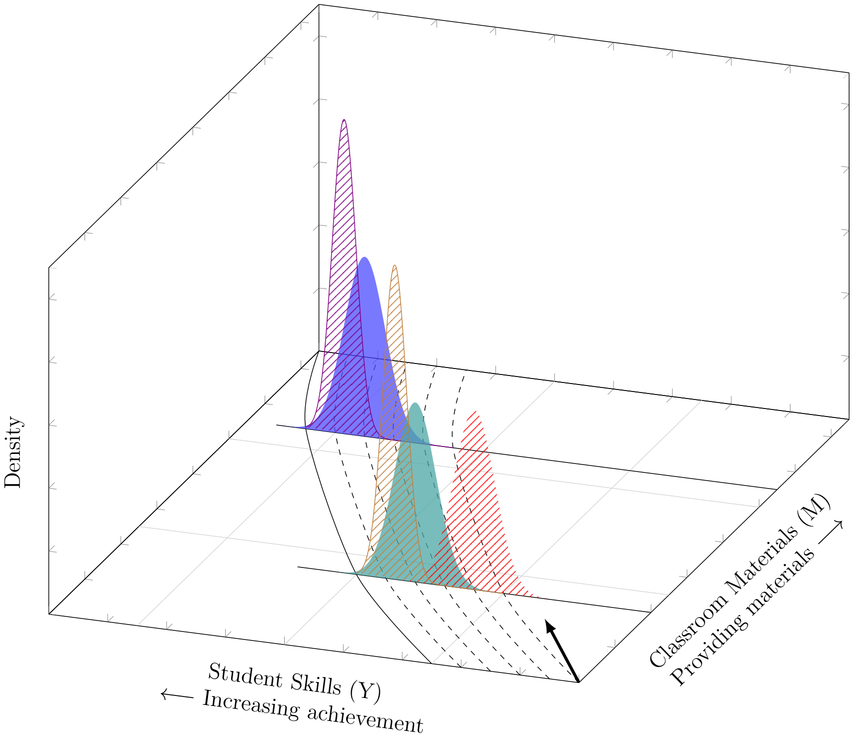
[12] Another potential driver of a positive or negative correlation across subjects is that subjects are complements or substitutes in terms of student ability—for example, learning math may make learning physics easier, or learning how to read may make learning how to write easier. We cannot differentiate these two channels.

2.4 Variation in Teacher Value-Added

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.[13] 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.

Figure 2

A Simple Model of Teacher Effectiveness and the Production of Learning



[13] 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.

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.

2.5 What Happens with a Teacher-Focused Intervention?

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 production 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.

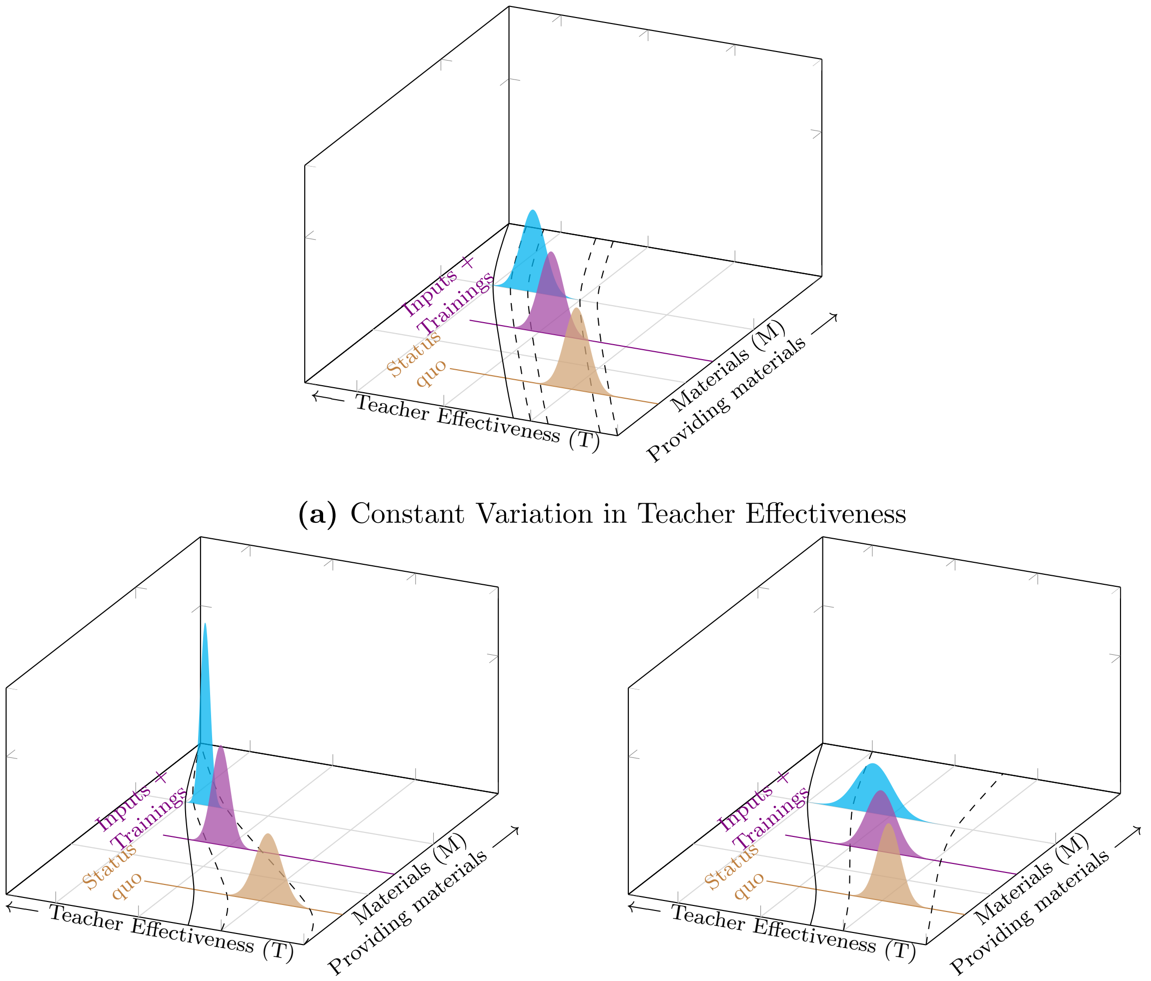
2.5.1 Rank Preservation

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

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

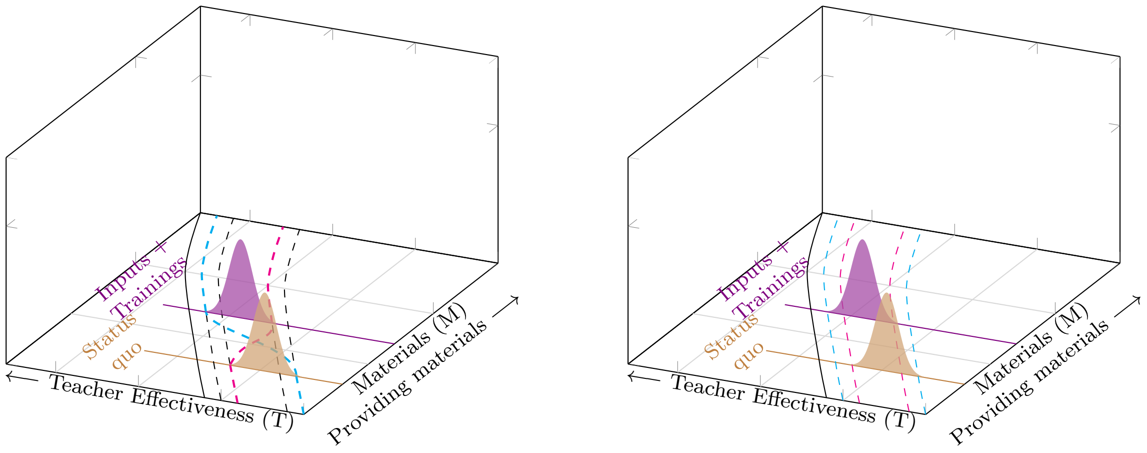
training provided ("skills beget skills") and thus benefit relatively more than low-performing teachers.

2.5.2 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, 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

Figure 4

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



(a) Crossing Productivity Curves

(b) Leap-Frogging Effective Teachers

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.

2.5.3 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 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 The NULP Evaluation and Sorting

This section provides the context for the study by describing primary schooling in Uganda, discussing the NULP intervention as well as the evaluation including describing the data and sample of schools, students and teachers, and lastly providing descriptive evidence of sorting of teachers and students.

3.1 Context - Primary School in Uganda

This section describes the setting of the study in Northern Uganda as well as an overview of primary schooling and teachers in Uganda.

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 its population is considered "poor", over twice the rate of any other region (Ministry of Finance 2023). 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 infrastructure, materials, or teacher support (Spreen and Knapczyk 2017). The constraints that we outline for sub-Saharan or Ugandan teachers more generally, are especially challenging for teachers in Northern Ugandan schools.

Primary education in Uganda consists of seven years of schooling, from grade one through grade seven.Ugandan students face major challenges to learning. Bold et al. (2017) find that the vast majority (94 percent) of children in government primary schools can not read even a simple paragraph. Among students in grade seven, 20 percent are unable to read and understand a short story (Uwezo 2016).[14]

In Uganda, there are 11 different languages of instruction. In 2007, the government mandated local-language instruction in the lower primary grades (grades one to three). However, there are many obstacles to implementing this "mother-tongue first" policy, including underdeveloped language orthographies, poorly trained teachers, and a lack of relevant 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 has been limited (Altinyelken, Moorcroft, and Draai 2014).

Primary school teachers in Uganda must obtain a teaching certificate requiring four years of secondary school and two years of pre-service teacher training; some teachers have additional qualifications, receiving a diploma. Pre-service training is generally of poor quality, with limited practical classroom experience (Hardman et al. 2011). Teachers in Uganda receive Continuous Professional Development (CPD), consisting of in-service training intended to update classroom competencies. The CPD program is managed through primary teachers' colleges by Coordinating Center Tutors (CCTs). CCTs are government employees assigned

[14] These statistics may even overstate student performance because schools discourage weaker students from attending grade seven to encourage the strongest students for the higher-stakes primary leaving exam (Gilligan et al. 2018).

to Coordinating Centers, similar to a school district in the United States. They are typically recruited from experienced teachers and head teachers (principals) and are responsible for providing in-service training workshops and school-based support such as conducting classroom observations and providing feedback to teachers and head teachers. CCTs, however, receive limited training, support, or financial resources, 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 under-trained, 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 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.

3.2 NULP Intervention and Evaluation

The Northern Uganda Literacy Project (NULP) was an early-grade mother-tongue literacy program developed in response to the educational challenges facing Northern Uganda. In this sub-section, we describe the details of the NULP intervention and evaluation that was conducted from 2013 to 2017, including randomization, how schools, students, and teachers were sampled, and the data.

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 of Northern Uganda, where the vast majority of the population speaks one language—Leblango. The NULP's approach involves directly training and supporting teachers with a new pedagogical approach as well as providing classroom and teacher educational materials. The program provided teachers three, week-long residential trainings on local-language orthography, as well as well as a slower-paced, phonics-based literacy curriculum. Teachers also received monthly classroom support visits from Mango Tree staff to provide feedback and support. Teachers were provided with detailed, scripted guides that outlined daily and weekly lesson plans. Each classroom was provided new textbooks and readers for students, and first-grade classrooms were provided slates, chalk, and wall clocks.

Over the course of the study, the NULP was introduced to different grades. Appendix Table A1, Panel A presents this information. In 2013 and 2014, only first-grade teachers in treatment schools received the NULP. In 2015, only second-grade teachers received the program, and in 2016, only third-grade teachers received the program. Classrooms were allowed to keep all of the Mango Tree educational materials in the year(s) after receiving the program, but teachers no longer received additional training or support visits.

We evaluate two versions of the NULP. The full-cost version consisted of the literacy program implemented directly by Mango Tree staff. A modified version, the reduced-cost NULP followed a "cascade" or "training-of-trainers" delivery model led by Ministry of Education CCTs rather than Mango Tree staff. The reduced-cost NULP also involved fewer classroom support visits. More details on the differences between the two versions is outlined in Kerwin and Thornton (2021).

3.2.2 Sample of Schools and School Random Assignment

The evaluation involves 128 schools, sampled in two phases. In 2013, 38 eligible schools were selected to be part of the study. Eligibility was established by Mango Tree, the most important criteria being that each school was required to have exactly two first-grade classrooms.[15] 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.[16]

Schools were randomly 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.[17] 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.

3.2.3 Sample of Students and Teachers

We follow four cohorts of first-grade children who entered the study schools in 2013, 2014, 2015, and 2016. Panel A of Appendix Table A2 describes the sampling strategy for students

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

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

[17] 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.

in the study, which varied by year. Students were either sampled at the baseline or at the endline.[18] Students were sampled stratified by gender and classroom.

Our sample of teachers to estimate value-added include any classroom teacher across the four cohorts of sampled students. This results in a sample of teachers of students in grades one to five. To shed light on sorting of teachers to grades, we also make use of school records for all teachers in grades one to seven.

3.2.4 Data

We use two types of data: measures of student learning–in local-language reading, English reading and Math, and characteristics of students and teachers.

Reading ability is measured using the Early Grade Reading Assessment (EGRA), an internationally recognized assessment of early literacy skills (Dubeck and Gove 2015; RTI 2009; Piper 2010; Gove and Wetterberg 2011). We use two different validated versions of the test—English and local-language reading. Both versions of the EGRA 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 EGRA also has a letter sounds module. Because both government regulations and the NULP curriculum stipulate that first-grade students only be exposed to local-language reading and writing, English reading assessments were conducted beginning in grade two. Students in first-grade were administered an oral English test to proxy for English reading ability. We conduct robustness analyses omitting observations using oral English as outcomes. Math ability is 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 at the grade-year-level, 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

[18] 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, with 60 students 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. In 2015, 30 first-grade students (Cohort 3) were randomly selected from each school at endline. 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.

group separately for each year and grade.

Test administration varied somewhat by subject, year, and grade. Appendix Table A1, Panel B summarizes the grades assessed each year, as well as the timing of the assessments. In 2013 and 2014, learning assessments were administered at the beginning and end of the school year (except for math), 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 three through five.

Throughout our analysis we include student-level controls for age, gender, and expected grade. In Buhl-Wiggers et al. (2024), we find a moderate treatment effect on grade progression. For this reason, we control for expected, or "on-track" grade as opposed to actual grade attended in a given year.

We also use teacher characteristics from teacher surveys and employee rosters. Teacher surveys were conducted in 2013 (grade 1 teachers), 2014 (grade 1 teachers), 2015 (grades 1-3), and 2017 (grades 3-5). Rosters of current and prior employees were collected from each school in 2014-2017. From these surveys and rosters, we have information on each teacher's age, gender, years of experience teaching, as well as years and level of education. Any time-varying variables (i.e. age and experience) are converted to their 2015 levels.

3.3 Sorting of Teachers and Students

Systematic 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.[19] In our data and empirical approach, we address three types of endogenous sorting: teachers to schools, teachers to grades (within schools), and teachers to students (i.e., students to classrooms within grades). In this subsection, we describe the extent of non-random sorting in the 42 control schools across the years of the study.

3.3.1 Sorting of teachers to schools

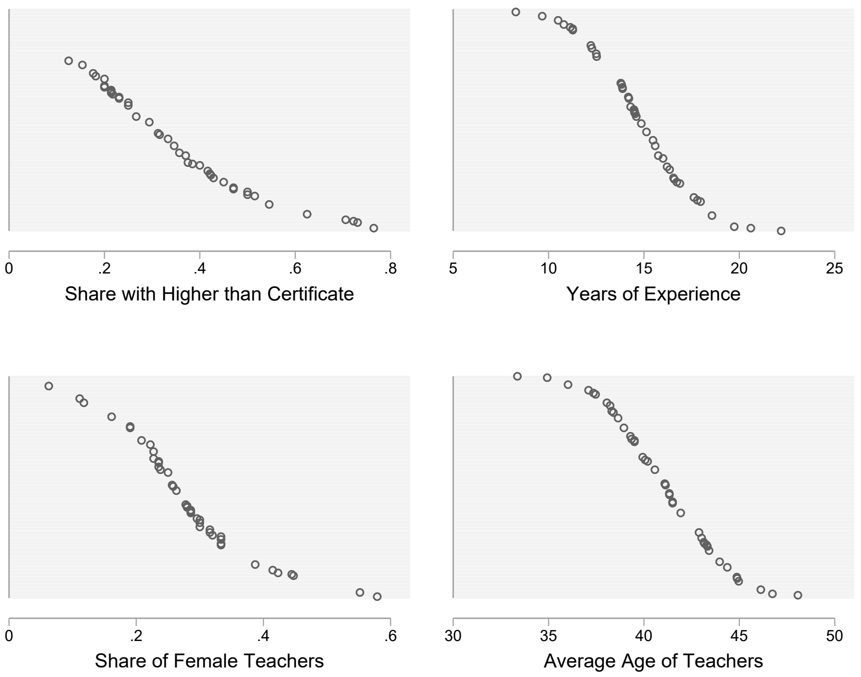
To understand the sorting of teachers to schools, we first graph the distribution of four teacher characteristics – having more than a teaching certificate, years of teaching experience, being female, and age – by school. We also plot these averages by geographical district, sorted by per-capita district level GDP (Rafa et al. 2017).

[19] We know of only a handful of papers. Glassow and Jerrim (2022) find evidence of sorting of teachers to schools according to experience using TIMSS data, Glassow, Franck, and Hansen (2023) do so using TALIS data, and Ajzenman et al. (2024) do so in Peru. Hannum (2001) discuss student sorting to teachers in Chinese schools.

Figure 5 shows the variation in teacher characteristics for the 42 control schools in our sample. The share of teachers within a school with more than a teaching certificate varies from 12 to almost 80 percent. The average percentage of female teachers within a school varies from 6 to almost 60 percent. There is also substantial variation in teacher age and years of teaching experience ranging between 35 to almost 50 years old and from an average of 7 to 23 years of experience. Together these results show substantial variation and that teachers are not equally distributed across schools.

Figure 5

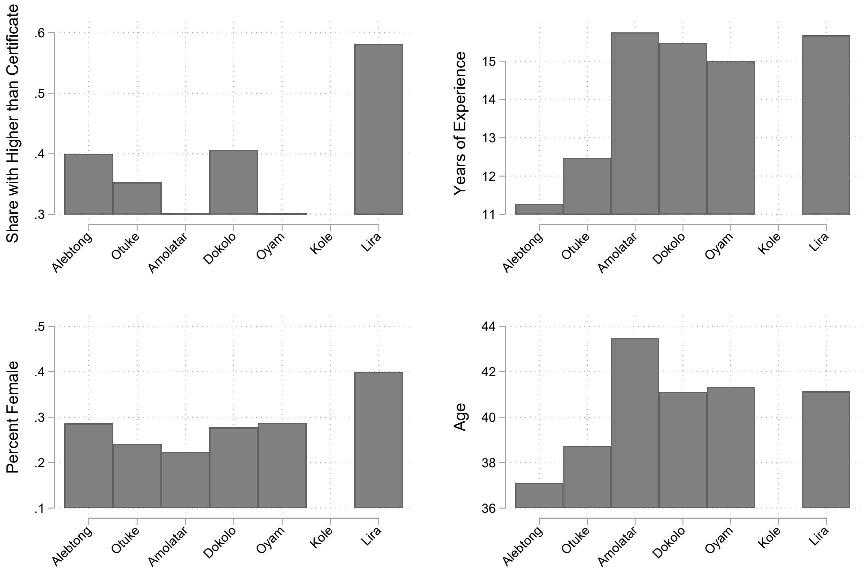
Teacher Characteristics by School



Notes: The panels in this figure graph the distribution of average teacher characteristics for 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, sorted by the district-level per-capita 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 (on the far left), are younger and have the lowest levels of experience. This pattern may reflect schools from poorer districts either being unable to recruit or retain experienced teachers.

Figure 6
Teacher Characteristics by District



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

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 sorting.

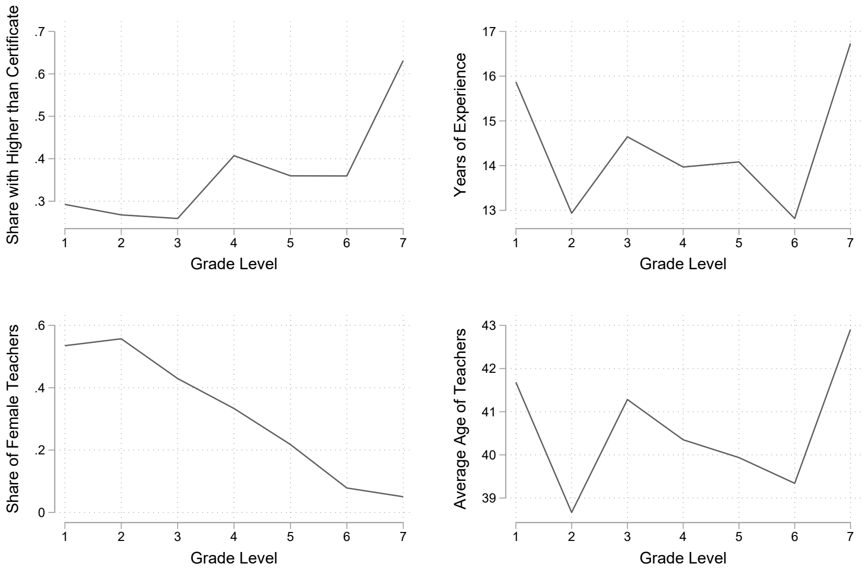
3.3.2 Sorting of teachers to grades

In Ugandan government primary schools, one teacher is typically assigned to one classroom, with multiple teachers teaching within a grade.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 plots the average teacher characteristics by grade. We see a u-shaped relationship with respect to years of teaching experience and age with grade, in which older and more experienced teachers are more likely to teach in grades 1 and 2 as well as in grades 6 and 7. The patterns in are consistent with more emphasis being placed on higher grades because of the importance of the grade seven primary leaving exam. Another explanation is that teachers with higher seniority may request to be placed in higher grades if those children are easier to teach, or if teaching those grades is more prestigious.

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.

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.

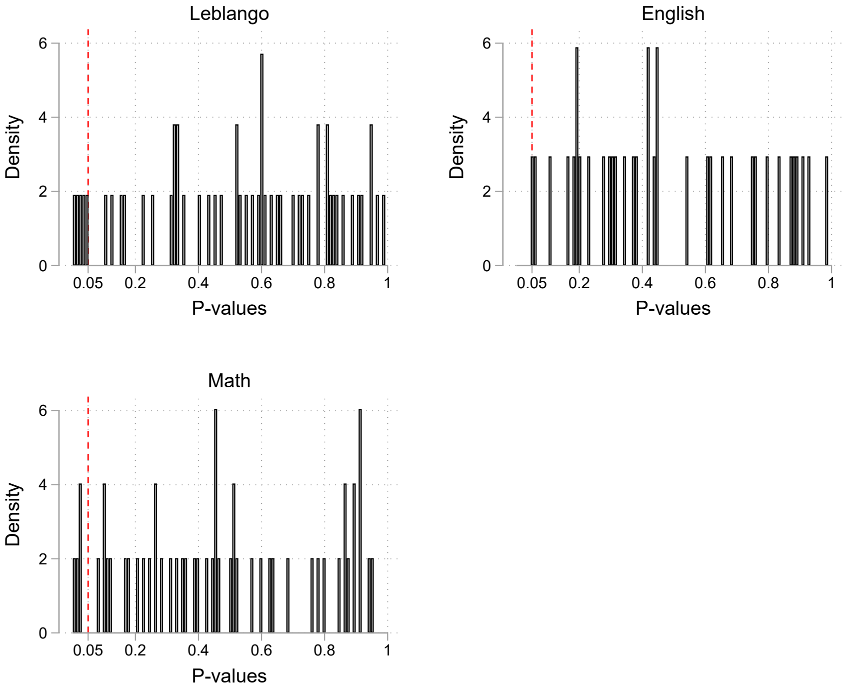
3.3.3 Sorting of students to classrooms within grades

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 (Rothstein 2009).

To investigate the degree of sorting of students to teachers/classrooms, we calculate the difference in student test scores between classes within-schools and grades each year within control schools, to test whether similar types of students sort into the same classroom following Horvath (2015). We present these results for 2014 and 2015, years in which, as we explain further below, schools received no instruction on how to sort students to classrooms. We find little evidence of student sorting: Figure 8 presents the p-values of the Horvath

(2015) tests indicating little evidence of bunching below the significance level of 5%.[20]

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, using data from 2014 and 2015 when schools were not instructed to randomize students to classrooms. The red vertical line mark a p-value of 0.05.

All in all, we find evidence of sorting of teachers to schools and grades but little evidence of sorting of students to classrooms within grades.

4 Analytical Samples and Descriptive Statistics

Prior to presenting the strategy for estimating classroom and teacher effects described in Section 5, in this section we describe our analytical samples of students and teachers, present descriptive statistics of the sample, and discuss balance and attrition.

[20] In local-language reading, 5 out of 53 comparisons (corresponding to 9%), show significant differences in baseline test scores at the 5 percent level.

4.1 Construction of Analytical Samples

In this subsection we describe the construction of our analytical samples to estimate classroom and teacher value-added.

4.1.1 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 assessment year.

For each student, we need an endline test score in 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.[21]

Because first grade students were not tested in English reading, we estimate English reading value-added only for students in grades two and above.[22] 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).

For some students, we have an endline test score but are missing a prior test score, if, for example, a student was absent on the day of an assessment. This occurs for approximately 10,000 student-year observations in each subject, corresponding to approximately 17% for local-language reading and math and 27% for English reading; the rate does not vary between treatment arms. In these cases, we impute students' missing prior test score to zero and provide robustness analysis omitting these observations. Appendix Table A3, Panel A presents the sample sizes of student-year observations with endline and consecutive-year tests for the entire sample, and separately by treatment arm.

4.1.2 Matching Students to Teachers

We match students to specific teachers using classroom registers and student reports.

[21] 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 grade one students who were assessed at the beginning of grade one, 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 for robustness, only impute scores that are missing and we show our results under these alternative specifications.

[22] 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.

Across 58,774 total student-year observations for which we have at least one endline test score in local-language reading, we can match approximately 99 percent to a teacher (Appendix Table A3, Panel B).[23]

To limit estimation error due to sampling variation, we drop student-year observations with fewer than five students per teacher in a given year (Appendix Table A3, Panel B). The rate of observations with fewer than 5 students is 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). This removes 2,442 student-year observations, or 4.2 percent of the overall sample, bringing us to 55,702 student-year observations for local-language reading.

4.1.3 Removing School and Grade Effects

To address sorting of teachers and students to schools and grades, the estimation strategy we pursue, outlined below, involves removing school or grade effects. To remove school effects, we need at least two teachers in each school. Because we follow the same schools over time, we can purge either overall school effects or year-specific school effects. We choose to purge overall school effects instead of year-specific school effects.[24] Similarly, to remove grade effects we need at least two teachers in each grade. Many grades only have one classroom in a given year. Therefore, to remove 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.

4.1.4 Teacher Effects vs. Classroom Effects

To separate teacher effects from classroom effects, we need to observe a teacher over multiple years. While we discuss teacher attrition in more detail below, we observe roughly 40 percent of the teachers teaching at least two years, generating the sample we use to calculate teacher effects.

[23] 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.

[24] Because the study design involved adding new cohorts of grade one students each year, we have fewer classrooms per school in earlier years of the intervention. This means that we have fewer teachers per school in earlier years. Purging year-specific school effects would result in dropping relatively more teachers from earlier years as we have more schools with only one teacher.

4.1.5 Final Analytical Samples

Table 1 shows the main analytical samples of schools, teachers, classrooms, and students that we use to estimate classroom effects (Columns 1 and 2) and teacher effects (Columns 3 and 4). We also show the number of student-year observations for each sample. This table presents the sample for control schools only, which provides our "status-quo" value-added estimates.

Within the classroom and teacher effects samples, Columns 1 and 3 show the samples used that remove school effects with at least two teachers in each school, and Columns 2 and 4 show the samples used to remove school-grade effects with at least two teachers in a given grade.

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
Students	8,814	7,940	7,624	6,260
Student-year obs	17,571	14,202	11,673	8,784

Notes: This table presents the analytical samples to estimate classroom and teacher value-added. The Classroom Effects sample involves having at least two teachers in each school (Column 1) or grade (Column 2). The Teachers Effects sample involves having at the same teacher over multiple years. The 42 control schools were sampled in two phases: 12 in 2013 and an additional 30 in 2014.

For both classroom and teacher effect estimates, we also perform analysis on a subset of teachers for whom we also have data on 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.

4.2 Descriptive Statistics, Balance, and Attrition

4.2.1 Descriptive Statistics and Balance

Appendix Table A4 presents descriptive statistics for students and teachers, separated by study arm. Half the students are female (by design) and are on average almost nine years old

(Panel A). Teachers are, on average, 40 years old, with 40-45 percent of them being female (Panel B). Teachers have 15 years of education and 14 years of experience. Most teachers have a teaching certificate, the minimum requirement for teaching in primary school, and around 30 percent have higher qualifications.

Column 4 presents the p-value from an F-test of means of each summary characteristic across the three study arms, showing that teacher and student characteristics are generally balanced across study arms.[25]

4.2.2 Attrition

Student attrition from the study could be due to dropping out, transferring to another school, or being absent for an assessment. The extent to which certain types of students attrit—either overall or differentially by study arm—could affect the external or internal validity of our analysis. Additionally, we might be concerned if student attrition is correlated with teacher ability.

There are several ways that we could define attrition. One measure of student attrition is being present for a baseline assessment but missing at the endline within the same year. 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.

To measure teacher attrition, we examine attrition of grade one teachers who were sampled in 2014 (when we have a larger sample of 128 schools) to 2016 (the last year that data was collected from grade one teachers). Of the 312 teachers in the data 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). Each year, as new teachers enter the sample, the likelihood of staying the next year is around 60 percent.

Appendix Table A7 presents the correlation between teacher characteristics and teacher attrition.[26] 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

[25] We regress each of the characteristics on a variable indicating study arm and cluster at the school level

[26] 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.

enter the sample are less likely to be female, less experienced and more educated, however this does not differ by treatment arm.[27]

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]

$$\begin{aligned} Y^k_{ijgs,t} =& \beta^k_1 Y_{ijgs,t-1} + \beta^k_2 Z_{ijgs,t-1} + \beta^k_3 X_{ijgs,t} + \lambda^k_{jgs,t} + \zeta^k_g + \beta^k_4 D_{ijgs,t} + \beta^k_5 ST_{ijgs,t} + \\ & \beta^k_6 Y_{ijgs,t-1}\zeta_g + \beta^k_7 Z_{ijgs,t-1}\zeta_g + \epsilon^k_{ijgs,t} \end{aligned} \tag{6}$$

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

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

[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.

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_6^k and β_7^k).

Our coefficients of interest are $\lambda_{jgs,t}^k$, 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 $C(M_{sjkt}, L_{sjkt}, E_{sjkt})$ 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 $\lambda_{jgs,t}^k$. We estimate $\lambda_{jgs,t}^k$ on our Classroom Effects sample, prior to restricting the data to the Teacher Effects sample.

When estimating Equation (6) we use both within- and between-school variation. This means that the estimate $\hat{\lambda}_{jgs,t}^k$ 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 $\hat{\lambda}_{jgs,t}^k$ 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):

$$\textit{within-school:} \qquad \hat{\gamma}_{jgst}^k = \hat{\lambda}_{jgst}^k - \frac{\sum_{c=1}^{C_s} N_{cs}\hat{\lambda}_{jgst}^k}{\sum_{c=1}^{C_s} N_{cs}} \tag{7}$$

$$\textit{within-grade:} \qquad \hat{\gamma}_{jgst}^k = \hat{\lambda}_{jgst}^k - \frac{\sum_{c=1}^{C_{sg}} N_{cs}\hat{\lambda}_{jgst}^k}{\sum_{c=1}^{C_{sg}} N_{cs}} \tag{8}$$

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 $\hat{\lambda}_{jgs,t}^k$ 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 (6), (7) and (8) 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:

$$\hat{\gamma}^k_{jgs,t} = \delta^k_{jgs} + \omega^k_{jgs,t} \tag{9}$$

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]

$$\hat{V}_{\text{corrected}}(\hat{\gamma}^k_{jgs,t}) = V(\hat{\gamma}^k_{jgs,t}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{[(\sum_{c=1}^{C_s} N_{cs}) - N_{cs}]}{N_{cs}(\sum_{c=1}^{C_s} N_{cs})}\hat{\sigma}^2\right\} \tag{10}$$

where $\hat{\sigma}^2$ is the variance of the estimated residuals, $\hat{\epsilon}^k_{ijgs,t}$, from Equation (6). 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:

[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.

$$\hat{V}_{\text{corrected}}(\hat{\gamma}^k_{jgs,t}) = V(\hat{\gamma}^k_{jgs,t}) - \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{[(\sum_{c=1}^{C_{sg}} N_{cs}) - N_{cs}]}{N_{cs}(\sum_{c=1}^{C_{sg}} N_{cs})}\hat{\sigma}^2\right\} \tag{11}$$

where we sum to the school-by-grade level (C_{sg}) instead of the overall school-level (C_s). $\hat{V}_{\text{corrected}}(\hat{\gamma}^k_{jgs,t})$ 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 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 3.3.3 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-

[33] For the correction of the variance of the teacher effects we use the following adjusted form of Equation (10): $\hat{V}_{\text{corrected}}(\hat{\theta}^k_{jgs}) = V(\hat{\theta}^k_{jgs}) - \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{[(\sum_{t=1}^{T_s} N_{ts})-N_{ts}]}{N_{ts}(\sum_{t=1}^{T_s} N_{ts})}\hat{\sigma}^2\right\}$, where $\hat{\sigma}^2$ is the variance of the residuals, $\hat{\epsilon}^k_{ijgst}$, from Equation (6). 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.

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, 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, for on control group schools only. We also provide various robustness and sensitivity analyses.

5.2.1 Variation in Classroom and Teacher Value-Added

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

Panel A presents the results that include both between- and within-school variation. 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 (Columns 1 and 2). Columns 3 and 4, as well as 5 and 6 present corresponding results for English and math, respectively. The results for English and math are somewhat larger with the estimated classroom and teacher effects for English at 0.52 and 0.43 SDs (Columns 3 and 4), and 0.51 and 0.42 SDs for math (Columns 5 and 6). Because these estimates include between-school variation, some proportion of the estimated variation is likely 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.

Panel B presents the within-school estimates, effectively comparing teachers between classes in the same school. Even after purging the estimates of school-specific effects, we still find substantial variation between teachers. For local-language reading, the estimated variance of classroom and teacher effects are slightly smaller than those in Panel A, at 0.33

[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.

[36] A full set of results including unadjusted estimates are presented in Appendix Table A11. We also compare adjusted to unadjusted estimates in Figure 9.

Table 2
Classroom and Teacher Value-Added: Control Schools

	Leblango		English		Math	
	Classroom (1)	**Teacher** (2)	**Classroom** (3)	**Teacher** (4)	**Classroom** (5)	**Teacher** (6)
Panel A: Between and within school variation						
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: Between grades and within school variation						
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: Between classes and within school and grade variation						
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: Between classes and within school and grade variation (Students randomly assigned)						
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.

SDs and 0.24 SDs, respectively. The variation in classroom and teacher effects for English in Panel B is very similar to the estimates for local-language reading, while the estimates for math are somewhat larger than local-language and English reading.

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 (Panel B) 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 primarily affects teacher effectiveness in English and math, where the decrease in estimates between Panel A to Panel B is larger than the decrease for local-language reading.

Panel C presents within school and grade estimates. Section 4 showed substantial sorting of teachers to grades within schools. Accordingly, Panel C shows within-grade estimates, resulting in rather large reductions (roughly two-thirds) in classroom and teacher effects for local-language reading. We also see large reductions in the estimates for English and math. 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.[37]

Lastly, Panel D presents the within school and grade estimates restricted to the years in which students were randomly assigned to classrooms. The change from Panel C is negligible, consistent with the results in Appendix Figure A1, suggesting limited systematic sorting of students to teachers.

Taken together, our preferred estimates are those in Panel C, as we take into account bias from non-random sorting of teachers to grades yet keep the largest possible sample. Figure 9 presents a summary of the estimates for local-language reading, visually plotting the different estimates. In Table 2 we only present results with the corrected SDs. However, Figure 9 shows each set of estimates with (in yellow) and without (in blue) adjustments for sampling error. The first three groups of estimates (from left to right) are classroom effects. The last two groups of estimates (from right to left) are teacher effects. The correction for sampling error reduces the estimated SD of effects by only around 10 percent for most of our estimates, and this adjustment only has a meaningful difference, across the two methods for within-school-grade effects (see Table A11).

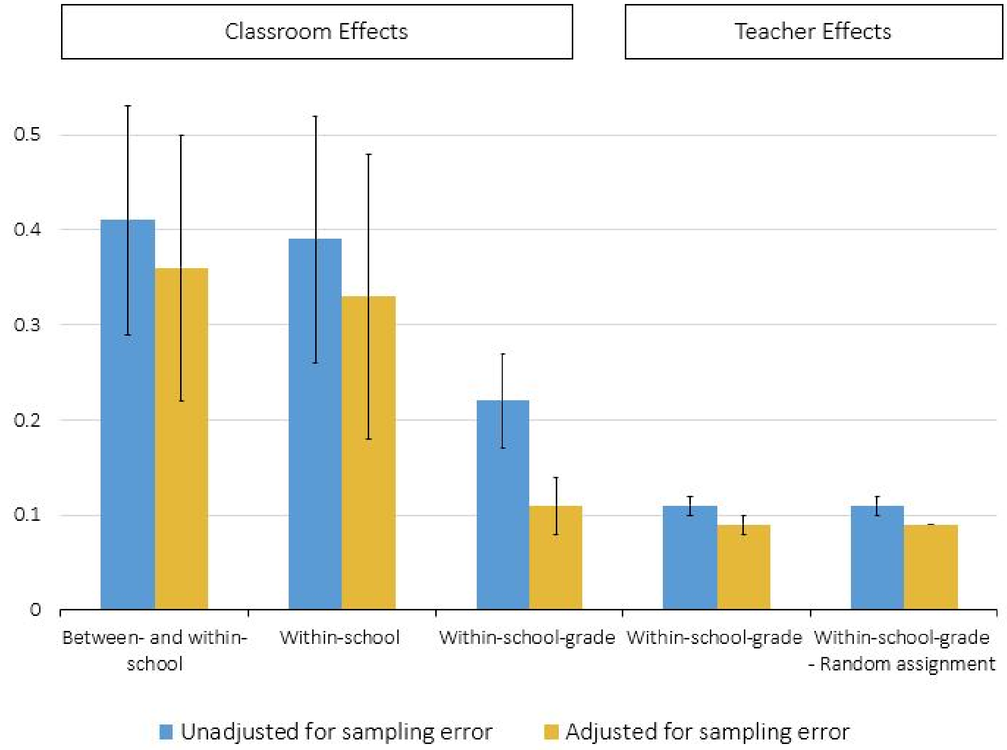
Classroom effects contains both the true effectiveness of a teacher as well as other time-specific effects such as peer dynamics or conditions on the day of testing. Thus, the difference in our estimated classroom and teacher effects shown in Table 2 and Figure 9 gives an indication of the fluctuation in classroom effects across years in our context. For local-language reading, the standard deviation of the year-to-year fluctuations are roughly two-thirds of the standard deviation of the teacher effects and even larger for both English and math. This suggests that there is a large variation in the year-specific classroom shocks in this context.[38]

Teacher value-added in local-language reading is positively correlated with English—the

[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 (correcting for sampling error) and Table A11 (not correcting for sampling error). The first three groups of estimates (from left to right) are classroom effects while the last two groups of estimates (from right to left) are teacher effects. The third group of estimates (within school and grade) show our preferred estimate of classroom effects, while the fourth group shows our preferred estimate of teacher effects.

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 0.16 with local-language reading. Looking at the correlation between classroom effects across subjects we see the same pattern. The correlation between local-language and English reading is 0.59, while the correlation between local-language reading is math is 0.22 and between English and math is 0.33. 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.

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) when student covariates or prior test scores are missing, and d) the use of alternative outcome measures.

We first address the changing sample composition of teachers and the potential implications for the estimates of classroom and teacher effects. First, only approximately 40 percent of the teachers used to calculate classroom effects are present across multiple years. This means that the sample of teachers used to calculate classroom and teacher effects in Columns 1 and 2 in Table 2 are different. Appendix Table A12 Panel A presents the classroom effects estimates using the same sample of teachers used to estimate teacher effects. The results are similar to those in Table 2. Second, 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 second set of robustness estimates relate to conditioning on classrooms with a minimum number of students. Our preferred estimates in Table 2 condition on having at least five students per teacher. Because 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 for these samples.

Third, we address the fact that the estimates in Table 2 involve imputing missing covariates—student age, gender, or prior year test scores. Appendix Table A14, Panel A presents the estimates without imputing age and gender, thus omitting any student-year observations with these missing covariates. The variances of the classroom and teacher effects differ only slightly from those in Table 2. Panel B of Appendix Table A14 presents the estimates omitting any student-year observations in which students do not have prior year test scores. This includes mainly first-grade students enrolled in the study in 2015 and 2016 because baseline tests were not administered in those years. The variances of the estimated classroom and teacher effects are only slightly different relative to those in Table 2.

Finally, we present sensitivity to using alternative outcomes. First, Appendix Table A15 Panel A shows the results for using use a "gain-score" model, in which rather than controlling for lagged test scores, we instead replace the left-hand-side of Equation (6) with $\Delta Y^{k}_{ijgs,t} =$

$Y^k_{ijgs,t} - Y^k_{ijgs,t-1}$. The results are similar to those in Table 2. Second, rather than combining all of the test score components using the first component of PCA. As robustness, Panel B of Table A15 presents the results using an alternative way of combining the test score components in which 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 changes 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 focused predominantly on local-language learning, we limit the results in this section to local-language reading outcomes. After comparing the distribution of value-added across NULP treatment arms including presenting sensitivity analysis, 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

This subsection presents the effect of the NULP on student learning, discusses how the NULP may have affected teacher quality, and presents our main empirical estimation strategy.

6.1.1 Effectiveness of the NULP on student learning

The NULP is a highly effective educational intervention. Appendix Table A16 shows the reduced form effects of being in a NULP treatment school on local-language reading, at the end of 2016. Column 1 presents the regression estimates being in a full-cost or reduced-cost treatment school, for students in cohort one, enrolled in grade one in 2014. Average student learning increased by 1.2 SDs in the full-cost version and 0.7 in the reduced-cost version for the this cohort of children after three years of exposure to the program. Our estimates of the effects of the NULP on measures of value-added pool across all available years of test score data, from 2014 to 2017. Column 2 presents the effects of the NULP across all four cohorts of students on their endline test score in 2014, 2015, 2016 or 2017, 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.[39]

6.1.2 Estimating the impact of the NULP on value-added

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.

Because the NULP rolled out in treatment schools to different grades across years (Appendix Table A1), teachers and students in some years and grades did not directly receive the NULP.[40] To maximize sample size, 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, which potentially provide a lower bound on the treatment effects. We also provide sensitivity analyses for grades and years for which students/teachers directly received the NULP in a given year.

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. to do so, 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.

6.1.3 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] Not all cohorts were equally or directly exposed to the NULP during these years (see Appendix Table A1). We present sensitivity analysis, below, of the effects of the NULP on teacher effectiveness only among students and teachers directly exposed to the NULP in the year of their learning assessment.

[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.

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:

$$Z^l_{dq} = \alpha^l_0 + \alpha^l_1 FullCost_s Q_q + \alpha^l_2 ReducedCost_s Q_q + \alpha^l_3 Q_q + \zeta^l_{strata} + \varepsilon^l_{dq} \qquad (12)$$

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 (12).

6.1.4 Correlation between value added and teacher characteristics

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

$$\hat{\delta}_{cgs} = \beta_0 + C'_{cgs}\beta_1 + \phi_s + \varepsilon_{cgs} \qquad (13)$$

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 and conduct sensitivity analyses. We then then test for rank preservation and correlate our value-added estimates with teacher characteristics to understand which teachers are most effective. 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 estimates for teachers in control group schools, identical to those 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.

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). Formally testing the difference across study arms, 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.

Table 3
Heterogeneity of Value-Added by NULP Study Arm

	Classroom Effects			Teacher Effects		
	Control (1)	Reduced-cost (2)	Full-cost (3)	Control (4)	Reduced-cost (5)	Full-cost (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

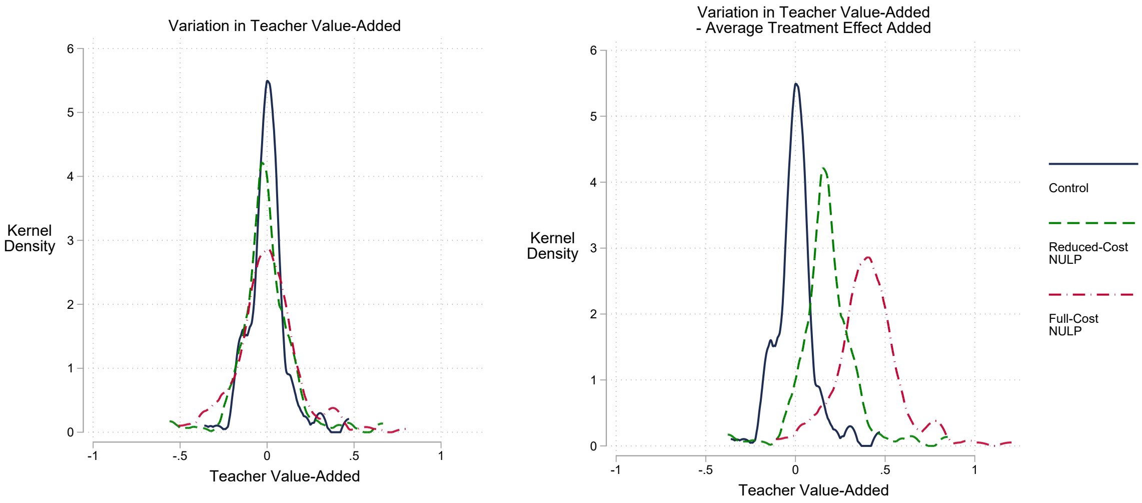
Notes: All estimates are purged of school-by-grade effects by subtracting off a weighted school-by-grade mean. 95% confidence intervals for the SD of the classroom effects are shown in brackets. The confidence intervals are cluster-bootstrapped using 1000 replications. To test the difference between 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.

In summary, two things happen in response to the NULP: 1) teachers, on average, become more effective (level shifts in student learning); and 2) teachers become more heterogeneous in their ability to affect learning (increased variance of teacher effectiveness). We illustrate these two effects in Figure 10.

The left hand side of Figure 10 presents the distribution of the teacher value-added from the within-school-grade estimates in Table 3, that purge out any level differences across grades and schools. These estimates also purge out the NULP treatment effect since the intervention took place at the school level.

The right hand side of Figure 10 illustrates the shift in levels and variation due to the NULP, on the distribution of value-added. To do so, we take the distribution of teacher value-added from Table 3 and manually add in the average NULP treatment effect from Appendix Table A16 to the full-cost and reduced-cost schools.

Figure 10
Treatment Effects on Teacher Value-Added



Notes: This figure presents the distribution of teacher effects from Table 3 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

We present three different sensitivity analyses for our results in Table 2 related to the effects of the NULP on the distribution of value-added. Our main focus for these additional analyses is based on the fact that the NULP intervention was only implemented in certain grade-levels across different years of the evalution (see Appendix Table A1). Across all the teacher in a NULP treatment school, more than half (60 percent) 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 three related sensitivity tests that account for the different levels of treatment exposure to teachers across grades and years.

First, the NULP was only fully implemented in treatment schools 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 were directly trained and supported by the NULP in the full-cost and reduced-cost treatment arms, and the corresponding students and teachers in the control group. 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 receiving training. These results are presented in Appendix Table A18, showing 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 who taught in higher grades with each subsequent year, thus being exposed to the NULP multiple times. Appendix Table A19 presents these results showing similar patterns to Table 3.

Finally, we 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 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 2.5.

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 2.5. 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

Table 4
Tests of Rank Preservation

	Age	**Female**	**Experience**	**Above Certificate**
	(1)	(2)	(3)	(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.

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.[41] 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.

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

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

Using five years of data from students and teachers in Northern Uganda, this paper presents the first estimates of the distribution of teacher effectiveness from sub-Saharan Africa. We also are the first to measure the causal effects of an educational intervention on the distribution of teacher effectiveness.

Our results highlight several key findings. First, despite the generally low learning levels in rural Uganda (and Africa more broadly), we find substantial variation in teacher effectiveness—some teachers are much more effective at increasing learning outcomes than others. We also show that sorting of students to teacher by ability is not an important concern for estimation in our setting. Instead, we provide evidence descriptively and in our estimation of value-added, that sorting of teachers to schools and grades is important. This has important implications for researchers using these methods with data spanning multiple grades. More descriptive work on classroom dynamics and sorting of teachers and students in Africa and beyond, may be helpful for understanding the implications of estimates of value-added across grades.

Second, the NULP caused both massive average gains in student learning and increased the variance of teacher effectiveness. Our theoretical framework predicts, and our results testing for rank preservation suggest, that the NULP most likely made 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.

Our findings contribute to the literature on teacher-value-added, emphasizing that comparing the variance of teacher effectiveness across different settings may not always be informative. Even in contexts where the best teachers have room for improvements, interventions can enhance teacher effectiveness at all levels of quality, or may only benefit teachers at the top (or only at the bottom). The variance in teacher value-added is usually interpreted as the scope for improving learning outcomes through teachers. Yet, we show that it is possible to increase both the average learning outcomes and the variance in teacher effectiveness through a targeted intervention, raising important questions about how to support lower-performing teachers and promote equity in teacher development programs. Static analyses of teacher effectiveness misses these type of insights.

Our results also indicate that observed teacher characteristics, such as education and experience, only explain a small fraction of the variance in teacher effectiveness. This underscores the limitations of *ex ante* screening of teachers based on these characteristics. Instead, more research is needed on how to design policies based on *ex post* evaluations and to identify alternative predictors of teacher effectiveness.

While our data come from Northern Uganda, many aspects of our setting–high enrollment, large classes, low learning levels, and frequent teacher absenteeism–are common across sub-Saharan Africa. Our findings suggest that even within such constrained environments, there is substantial variation in teacher value-added, offering meaningful scope for improving teacher effectiveness. Future research should investigate how the distribution of teacher value-added differs across settings in Africa, and examine how other educational interventions affect teacher effectiveness.

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.

Lastly, 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, our paper presents the first estimates of how the distribution actually shifts, as a causal effect of a teacher intervention.

References

Ajzenman, Nicolás, Eleonora Bertoni, Gregory Elacqua, Luana Marotta, and Carolina Méndez Vargas (2024). “Altruism or Money? Reducing Teacher Sorting Using Behavioral Strategies in Peru”. *Journal of Labor Economics* 42.4. Publisher: The University of Chicago Press, pp. 1049–1091. ISSN: 0734-306X. DOI: 10.1086/725166.

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.

Altinyelken, Hülya Kosar, Sarah Moorcroft, and Hilde van der Draai (2014). “The dilemmas and complexities of implementing language-in-education policies: Perspectives from urban and rural contexts in Uganda”. *International Journal of Educational Development* 36, pp. 90–99. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2013.11.001.

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.

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.

Buhl-Wiggers, Julie, Jason T. Kerwin, Ricardo Montero de la Piedra, Jeffrey A Smith, and Rebecca Thornton (2024). *Reading for Life: Lasting Impacts of a Literacy Intervention in Uganda*. 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”. en. *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 II: Teacher Value-Added and Student Outcomes in Adulthood”. en. *American Economic Review* 104.9, pp. 2633–2679. ISSN: 0002-8282. DOI: 10.1257/aer.104.9.2633.

Condie, Scott, Lars Lefgren, and David Sims (2014). "Teacher heterogeneity, value-added and education policy". en. *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". en. *Development Policy Review* 37.4, pp. 470–485. ISSN: 1467-7679. DOI: 10.1111/dpr.12371.

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*. Tech. rep. National Bureau of Economic Research.

Glassow, Leah Natasha, Emilie Franck, and Kajsa Yang Hansen (2023). "Institutional characteristics moderating the relationship between classroom socioeconomic composition and teacher qualifications: Evidence from 46 education systems in TALIS 2018". *International Journal of Educational Research* 119, p. 102170. ISSN: 0883-0355. DOI: 10.1016/j.ijer.2023.102170.

Glassow, Leah Natasha and John Jerrim (2022). "Is inequitable teacher sorting on the rise? Cross-national evidence from 20 years of TIMSS". *Large-scale Assessments in Education* 10.1, p. 6. ISSN: 2196-0739. DOI: 10.1186/s40536-022-00125-9.

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*. en. RTI International. ISBN: 978-1-934831-08-3.

Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge (2015). "An Evaluation of Empirical Bayes's Estimation of Value-Added

Teacher Performance Measures". en. *Journal of Educational and Behavioral Statistics* 40.2, pp. 190–222. ISSN: 1076-9986, 1935-1054. DOI: 10.3102/1076998615574771.

Hannum, Emily (2001). "Do teachers affect learning in developing countries? Evidence from matched student-teacher data from China". *conference "Rethinking Social Science Research on the Developing World in the 21st Century," Park City, Utah, June.*

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.* Tech. rep.

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.

Kinsler, Josh (2012). "Assessing Rothstein's critique of teacher value-added models". en. *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?" en. *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.* Tech. rep. Research Triangle Institute.

Rafa, Mickey, Jonathan D. Moyer, Xuantong Wang, and Paul Sutton (2017). "Estimating District GDP in Uganda". en. *SSRN Electronic Journal.* ISSN: 1556-5068. DOI: 10.2139/ssrn.3941446.

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 (2009). “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables”. *Education Finance and Policy* 4.4, pp. 537–571. ISSN: 1557-3060. DOI: 10.1162/edfp.2009.4.4.537.

— (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*. Tech. rep. 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”. en. *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.

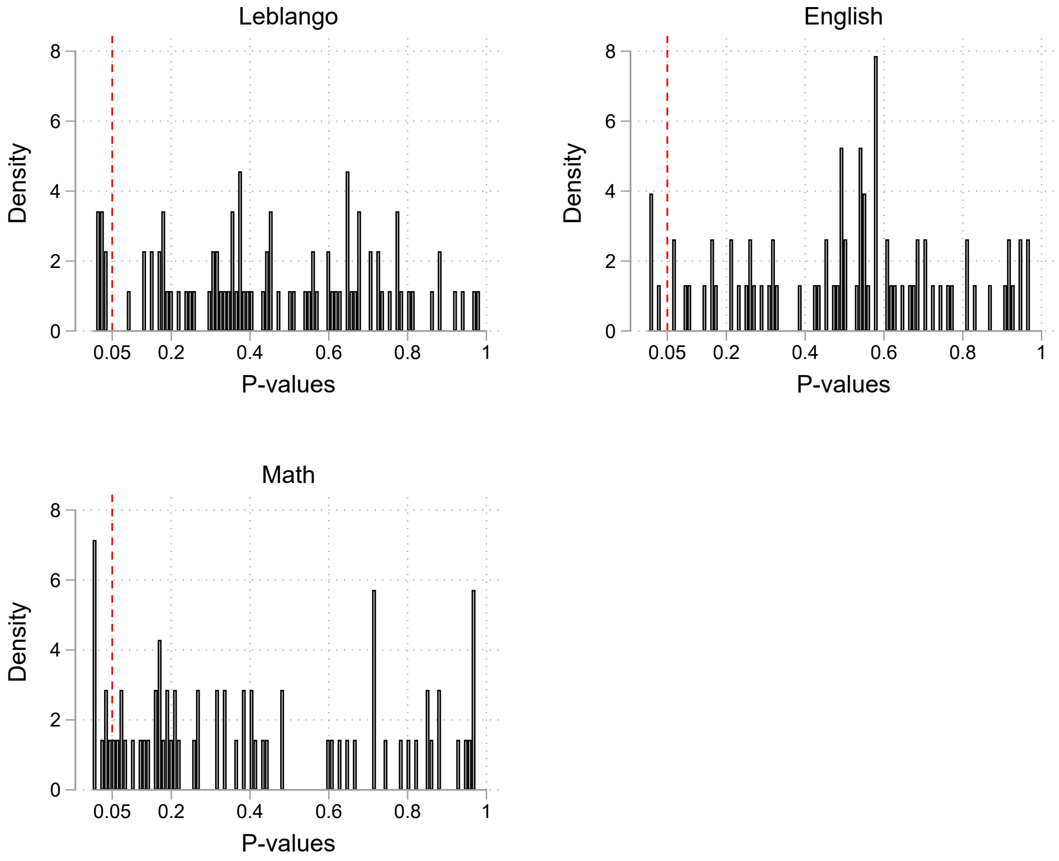
Todd, Petra E. and Kenneth I. Wolpin (2003). “On the Specification and Estimation of the Production Function for Cognitive Achievement”. en. *The Economic Journal* 113.485, F3–F33. ISSN: 1468-0297. DOI: 10.1111/1468-0297.00097.

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

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

	2013	**2014**	**2015**	**2016**	**2017**
Panel A: NULP Treatment	(1)	(2)	(3)	(4)	(5)
Grade receiving NULP	Grade 1	Grade 1	Grade 2	Grade 3	Grade 4
Panel B: 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
Panel C: Student Assignment to Classrooms					
Random assignment of students to classrooms	Yes	No	No	Yes	Yes

Appendix Table A2

Number of Students per School Sampled by School Sample and Year

	2013	**2014**	**2015**	**2016**
Panel A: Original 38 schools sampled in 2013				
Cohort 1 (Baseline sample)	50 grade-1 students			
Cohort 1 (Endline sample)		30 grade-2 students		
Cohort 2 (Baseline sample)		40 grade-1 students		
Cohort 2 (Endline sample)		60 grade-1 students		
Cohort 3 (Baseline sample)			30 grade-1 students	
Cohort 3 (Endline sample)				30 grade-2 students
Cohort 4				60 grade-1 students
Panel B: New 90 schools sampled in 2014				
Cohort 2 (Baseline sample)		80 grade-1 students		
Cohort 2 (Endline sample)		20 grade-1 students		
Cohort 3 (Baseline sample)			30 grade-1 students	
Cohort 3 (Endline sample)				30 grade-2 students
Cohort 4				60 grade-1 students

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

Appendix Table A3

NULP Sample Across Study Arms

	All	Control	Reduced Cost	Full Cost
Panel A: Students with Consecutive Tests				
Student-year obs with endline Leblango 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 B: 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 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. We regress each of the characteristics on a variable indicating study arm and cluster at the school level.

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

	Control	**Reduced-cost**	**Full-cost**	**All**
Teacher characteristics	(1)	(2)	(3)	(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

	Control	**Reduced-cost**	**Full-cost**	**All**
Teacher characteristics	(1)	(2)	(3)	(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

	Control	**Reduced-cost**	**Full-cost**	**All**
Teacher characteristics	(1)	(2)	(3)	(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 (1)	**Teacher** (2)	**Classroom** (3)	**Teacher** (4)	**Classroom** (5)	**Teacher** (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 (1)	Reduced-Cost (2)	Full-Cost (3)	Control (4)	Reduced-Cost (5)	Full-Cost (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 (1)	Reduced-Cost (2)	Full-Cost (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.