ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Buhl-Wiggers, Julie; Kerwin, Jason; Muñoz, Juan Sebastián; Smith, Jeffrey A.; Thornton, Rebecca L.

Working Paper

Some Children Left Behind: Variation in the Effects of an Educational Intervention

IZA Discussion Papers, No. 13598

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Buhl-Wiggers, Julie; Kerwin, Jason; Muñoz, Juan Sebastián; Smith, Jeffrey A.; Thornton, Rebecca L. (2020) : Some Children Left Behind: Variation in the Effects of an Educational Intervention, IZA Discussion Papers, No. 13598, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/224040

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13598

Some Children Left Behind: Variation in the Effects of an Educational Intervention

Julie Buhl-Wiggers
Jason Kerwin
Juan Sebastián Muñoz
Jeffrey Smith
Rebecca Thornton

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13598

Some Children Left Behind: Variation in the Effects of an Educational Intervention

Julie Buhl-Wiggers
Copenhagen Business School

Jason Kerwin
University of Minnesota

Juan Sebastián Muñoz
IÉSEG School of Management

Jeffrey Smith
University of Wisconsin and IZA

Rebecca Thornton
University of Illinois

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Some Children Left Behind: Variation in the Effects of an Educational Intervention*

We document substantial variation in the effects of a highly-effective literacy program in northern Uganda. The program increases test scores by 1.4 SDs on average, but standard statistical bounds show that the impact standard deviation exceeds 1.0 SD. This implies that the variation in effects across our students is wider than the spread of mean effects across all randomized evaluations of developing country education interventions in the literature. This very effective program does indeed leave some students behind. At the same time, we do not learn much from our analyses that attempt to determine which students benefit more or less from the program. We reject rank preservation, and the weaker assumption of stochastic increasingness leaves wide bounds on quantile-specific average treatment effects. Neither conventional nor machine-learning approaches to estimating systematic heterogeneity capture more than a small fraction of the variation in impacts given our available candidate moderators.

JEL Classification: I25, I26

Keywords: essential heterogeneity, heterogeneous treatment effects, education

Corresponding author:
Jeffrey Smith
Department of Economics
University of Wisconsin
Sewell Social Science Building
1180 Observatory Drive
Madison, WI 53706, 608-262-3066
USA

E-mail: econjeff@ssc.wisc.edu

* We thank seminar participants at the Heckman 75th Birthday Conference, Paul Glewwe, and Lois Miller for helpful comments, Brigham Frandsen for assistance in implementing the Frandsen-Lefgren bounds, and Tanya Byker and Sebastian Calónico for sharing their Stata code for testing rank preservation. Deborah Amuka, Victoria Brown, and Katie Pollman of Ichuli Institute were indispensable to the data collection for this study. This project would not have been possible without the efforts of Anne Alum, Patrick Engola, Craig Esbeck, Jimmy Mwoci, James Odongo, JB Opeto, and the rest of the Mango Tree Uganda staff, who developed and carried out the NULP intervention. We also thank the students, parents, and teachers from our study schools in northern Uganda. The usual disclaimer applies.

1 Introduction

This paper examines treatment effect heterogeneity in the context of an educational intervention implemented in northern Uganda. Kerwin and Thornton (2020) and Buhl-Wiggers et al. (2018a) show that the intervention—the Northern Uganda Literacy Project (NULP)—has an extraordinarily large average treatment effect, e.g. relative to the education interventions reviewed in meta-analyses (McEwan 2015; Evans and Yuan 2019). Yet the underlying student-level outcomes show that many students in the treatment group continue to have very low test scores even after multiple years of exposure. This observation motivates our title and suggests the presence of meaningful effect heterogeneity. Studying—and understanding—treatment effect heterogeneity can shed important light on how interventions work, for whom they work, and how they affect inequality.

There is a broad concern about some students being "left behind" and failing to learn (Rudalevige 2003). This issue was highlighted in the United States by the No Child Left Behind Act of 2002, but looms even larger in developing countries. A recent World Development Report focused solely on the "learning crisis" in developing countries (World Bank 2018): while school enrollment rates have risen substantially, many students learn almost nothing in school (Boone et al. 2013; Piper 2010). Similarly, Goal 4 of the Sustainable Development Goals addresses equity and inclusiveness in education, and UNESCO has emphasized that "every learner matters and matters equally" (UNESCO 2017).

Our analysis proceeds in three stages. We first establish that meaningful treatment effect heterogeneity exists by using the classical statistical bounds due to Fréchet (1951) and Höffding (1940) to bound the variance of the treatment effects [henceforth the "FH bounds"]. Related formal statistical tests indicate that we can clearly reject the common effect model. The second stage of our analysis considers what we can learn about effect heterogeneity by imposing additional assumptions. We first impose the property of "mutual stochastic increasingness" on the joint distribution of treated and untreated outcomes. This allows us to calculate the bounds on the average treatment effects at particular quantiles of the outcome distributions described in Frandsen and Lefgren [hereinafter "FL"] (2020). We then impose (and test) the stronger property of rank preservation in the context of an analysis of quantile treatment effects, which also informs us regarding the effects of the NULP on educational inequality.[1]

In the third stage of our analysis, we address the question of how the intervention works (and for whom) by looking for moderators—variables not affected by treatment that capture

[1] The literature offers a variety of other substantive assumptions that aim to reduce the identified set of treatment effect distributions; see, e.g. Bhattacharya, Shaikh, and Vytlacil (2008). We leave these for future work.

meaningful variation in the treatment effect. Following Djebbari and Smith (2008) we can think of our analysis as dividing the extant treatment effect variation into "systematic" (captured by the moderators) and "idiosyncratic" (not captured by the moderators) components. As they note, our ability to capture systematic heterogeneity depends crucially on the set of available candidate moderators. We conduct our search for meaningful moderators using both a traditional approach of looking for first-order interactions between the treatment indicators and various "usual suspects" and via the machine learning algorithm laid out in Knaus, Lechner, and Strittmatter (2020).

Our core finding is that the effects of the program vary widely across individual students. The lower bound on the standard deviation of treatment effects exceeds one standard deviation. This means that despite a massive average gain of 1.4 standard deviations, a normal distribution of treatment effects implies that the intervention harms over 8% of students, while a similar share experience individual gains in excess of 3.0 SDs. As a point of comparison, we show that the average effect of a reduced-cost version of the same program equals 0.7 SDs.[2] Thus the range of treatment effects for students within a given version of the program is over four times the difference in average effects. It also exceeds the gap in the average effects between the most-harmful and most-beneficial interventions reviewed in McEwan (2015).[3]

At the same time, our analyses make little headway in organizing the treatment effect heterogeneity the data clearly contain. The FL bounds do provide some insight and suggest that negative average treatment effects occur at the top of the outcome distribution if they occur at all. Rank preservation could provide a tight characterization of the heterogeneity but we easily reject its implications in our data. Our conventional moderation analyses explain essentially none of the variation in treatment effects. Even machine-learning methods using our available covariates do not help much: subtracting off the estimated conditional average treatment effects reduces the lower bound on the impact standard deviation by less than five percent.

Our findings imply that the extensive literature documenting the average effects of education interventions is fundamentally insufficient to address the ongoing learning crisis in developing countries. This massive literature has generated almost no information about how the effects of individual interventions vary across students. Eight recent papers have reviewed this evidence on "what works" in education in developing countries.[4] These studies

[2] The reduced-cost version reduces costs by about 64 percent.

[3] We use the McEwan review instead of one of the seven other systematic reviews because he helpfully made the underlying data on individual studies' treatment effects publicly available.

[4] The reviews are Glewwe et al. (2013), Kremer, Brannen, and Glennerster (2013), Krishnaratne, White, and Carpenter (2013), Murnane and Ganimian (2014), McEwan (2015), Evans and Popova (2016), Glewwe

collectively cover hundreds of randomized trials in dozens of countries. But they focus almost entirely on average treatment effects. None of them consistently report or discuss treatment effect heterogeneity, or emphasize the fact that individual treatment effects could deviate drastically from the average effect.[5] This pattern likely reflects the underlying literature: papers rarely report measures of treatment effect variation.[6] Even re-analyses of the raw data may yield limited evidence, since studies are commonly powered to detect only average effects, and do not have the sample sizes needed for measuring treatment effect heterogeneity (Glewwe and Muralidharan 2016).

Our paper offers several contributions to the existing literature. Substantively, we do a "deep dive" into treatment effect heterogeneity in a very different context than earlier efforts by Heckman, Smith, and Clements [hereinafter "HSC"] (1997), Djebbari and Smith (2008), and Bitler, Gelbach, and Hoynes (2017). Not only does northern Uganda differ greatly from the United States or rural Mexico, but the NULP educational intervention we study differs greatly from the active labor market program considered in HSC (1997), the PROGRESA conditional cash transfer program considered by Djebbari and Smith (2008), and the welfare-to-work program considered by Bitler, Gelbach, and Hoynes (2017). Our substantive findings regarding the clear presence of what Heckman, Urzua, and Vytlacil (2006) dubbed "essential heterogeneity" combined with our general failure to systematize that heterogeneity via observed moderators (even when applying machine learning methods) defines a clear path forward for future evaluations of educational interventions: empiricists should collect improved candidate moderators and applied theorists should devote themselves to motivating new moderators.

Methodologically, our paper represents only the second empirical application of the FL (2020) bounds and arguably the first with a data set of meaningful size. In addition, while numerous recent papers examine treatment effect heterogeneity using one or the other of the vast array of competing machine learning algorithms currently in circulation, we add value by comparing traditional a priori methods to one particular machine learning algorithm. We do this within the context of a broader discussion of theories of treatment effect moderation, and of the relative values (at the margin, of course) of creative data collection and the further

and Muralidharan (2016) and Conn (2017)

[5] Glewwe and Muralidharan (2016) point out that treatment effect heterogeneity is "likely to be a first-order" issue, but that standard practice focuses on average effects. Four of the reviews discuss analyses of systematic heterogeneity in the effects of one specific intervention: Murnane and Ganimian (2014), Evans and Popova (2016), Glewwe and Muralidharan (2016) and Conn (2017). Interestingly, each chooses to highlight a different intervention for this purpose.

[6] Evans and Yuan (2018) review 281 evaluations with learning outcomes, conducted between 2000 and 2016; only 33 percent presented results for boys and girls, 23 differentiated effects by baseline achievement, and only 11 percent differentiated effects by socio-economic status.

refinement of what are often already quite esoteric statistical moderator selection schemes. We also show that even when machine learning techniques identify important variation in treatment effects, they can still leave a large amount of treatment effect heterogeneity unexplained. This finding has a substantive implication: papers that use these techniques should report bounds on the impact variance before and after removing the estimated systematic heterogeneity.

Our findings imply that doing much better at predicting treatment effects will mean going beyond the potential moderators typically available to schools or educational authorities in their administrative data, particularly in developing countries. If researchers collect better moderators, they could use them to alter the design of programs in order to trim the lower tail of treatment effects while holding steady, or even increasing, the average gains. At the same time, moderators not immediately available to administrators will add less in an immediate, practical sense.

The remainder of the paper takes a familiar course. We begin by describing the intervention we study in Section 2 and by describing the data we analyze in Section 3; Section 4 presents the average treatment effects of the program for reference. Section 5 covers the FH bounds, and establishes the presence of treatment effect heterogeneity. Section 6 tries to reduce this heterogeneity by imposing additional assumptions, first mutual stochastic increasingness and then rank preservation.. Section 7 documents our search for meaningful moderators, first using the traditional a priori approach and then using (one particular) machine learning strategy. Finally, Section 8 reviews our results and ties them back into the broader literature.

2 Northern Uganda Literacy Project (NULP)

2.1 Context

Our study takes place in the Lango sub-region of northern Uganda, one of the poorest regions of the country. The primary education system in northern Uganda faces major challenges.[7] The pupil-to-teacher ratio exceeds 65:1 and the pupil-to-classroom ratio is about 86:1. On an average day about 28% of teachers and 24% of students miss school (Bold et al. 2017; Ministry of Education and Sports 2016; Uwezo 2019). The vast majority of schools in our data lack electricity, though nearly all have at least one latrine. Poor schools, particularly when combined with a history of civil conflict, lead to poor outcomes: the adult literacy rate

[7] Primary education in Uganda runs from P1 (first grade) to P7 (seventh grade), with students typically entering P1 at age six.

in the Lango sub-region sits just above 71 percent, with even lower rates in other sub-regions of northern Uganda (Uganda Bureau of Statistics 2017).

Until fairly recently, the curriculum and pedagogy of the schools we study reflected Uganda's British colonial past. Indeed, today's classrooms—often entirely in English with a call-and-response pedagogy—sound a bit like one would imagine a working-class school in Manchester or Birmingham in the first half of the last century. In 2007, the government of Uganda implemented a new primary education curriculum aimed at improving on this history. This new curriculum remains in place today and includes two important features. First, students in P1-P3 must be taught in the main mother tongue of their area with a transition in P4 to full English instruction starting in P5. Second, teachers devote an hour to literacy lessons each day, with the first half hour on reading and the second on writing.[8] In practice, many teachers had trouble adjusting to the new curriculum due to limited access to materials, underdeveloped orthographies of local languages, and inadequate training, so these policies remain only partially implemented; see e.g. Altinyelken (2010) or Ssentanda (2013).

Schools in Uganda also face shortages of teaching materials. The central government provides an annual allocation to each school for supplies that varies with enrollment and with the government budget and typically lies in the range USD$1000-$2000. Per Kayabwe et al. (2014), schools must allocate these funds exclusively into four categories—teaching and learning materials, extra-curricular activities, school management, and school administration—but have discretion over the share allocated to each category. In addition, the Enhancement of Universal Primary Education and Community (EUPEC) program provides supplementary learning materials.[9] Unfortunately, the annual allocation and the EUPEC, combined with "contributions" from parents, still leave the schools we study in dire need of instructional materials.[10]

Overall, despite the various reforms, the quality of education in Uganda remains low. According to Piper (2010), 80% of students in the Lango sub-region could not read a single word of Leblango at the end of P2 and 50% could not at the end of P3. Similarly, Bold et al. (2017) find that 42% of students in Uganda could not read a word at the end of P3.

[8] Perhaps surprisingly, this was not happening prior to the reform; see, e.g. Read and Enyutu (2005).

[9] For example, EUPEC mandated the distribution of slates to all P1 and P2 classrooms, but many of the slates ended up in higher grade classrooms or lost due to poor maintenance.

[10] Primary school is fee-free in Uganda, but schools almost always ask parents to make monetary contributions. Though nominally voluntary, parents who do not contribute may find their children sent home from school. In our control group, 97 percent of parents report making contributions with a median of about US$9. Among other things, the money is used to pay for teacher bonuses and housing, which the central government does not cover.

2.2 Intervention

From 2009 to 2013, Mango Tree—a private, for-profit, educational tools company—developed a model program designed to address the main challenges in primary education in northern Uganda. The program, called the Northern Uganda Literacy Project (NULP), focused on mother tongue literacy in P1 to P3 by training and supporting teachers with additional classroom materials and a revised pedagogy.

The program consisted of four main features. First, it provided teachers with intensive training in teaching mother tongue literacy. This component of the program included residential training sessions as well as in-class supervision visits that provided teachers with feedback.[11] Second, the NULP provided teachers with classroom materials that supported their training. Program classrooms received primers (textbooks that follow the curriculum), readers (books for reading practice), teacher guides that describe in detail the specific lesson plans for each day, slates that allow students to practice their writing[12], and a wall clock used for monitoring time during lessons. Third, the NULP model followed the government curriculum in teaching in students' mother tongue in P1 and P2, but introduced letters and sounds at about half the usual pace—covering the first half of sounds in P1, with the second half in P2. Oral English was introduced as subject in P1, and then added into lessons in P2 and P3, to allow time for students to develop critical early literacy skills before pushing them to use those skills in another language.[13] Finally, the NULP model engaged with parents by promoting the benefits of mother-tongue instruction using a radio program, and holding school meetings to train parents on how to support their children's learning at home.[14] According to Kerwin and Thornton (2020), the marginal cost of NULP equals about $20 per student per year, relative to a base level of expenditures in Ugandan primary schools of around $60 per student per year.

Package interventions like the NULP have received little attention from researchers in the developed world. In developing-country contexts, in contrast, a number of interventions

[11] Training sessions were distributed over the school year and included three residential trainings during school holidays and six in-service training workshops on Saturdays. Trainers used a detailed facilitator's guide as well as instructional videos. Supervision visits were carried out by Mango Tree staff, coordinating center tutors who work for the Ministry of Education, and mentor teachers who had previously experience teaching the NULP instruction model.

[12] The slates are not, as readers in developed countries might surmise, electronic tablets or e-readers. Rather, a slate resembles a small piece of a blackboard that can be used with a piece of chalk.

[13] Some research suggests that teaching in a student's mother tongue motivates and enhances academic achievement. Students instructed in a familiar language gain early reading skills more quickly and exhibit improved attendance. Moreover, the evidence suggests that the use of native languages in the first years of education enhances the learning of a second tongue in later years (Webley 2006). Smits, Huisman, and Kruijff (2008) review this literature.

[14] In these meetings parents learn the importance of mother-tongue instruction, as well as how to assess and support children's learning and literacy development at home.

provide combinations of inputs similar to that in the NULP. Examples include the Primary Math and Reading (PRIMR) Initiative in Kenya (Piper, Zuilkowski, and Ong'ele 2016) and the School Health and Reading Program (SHRP) in Uganda (Brunette et al. 2019). Other interventions provide some of the inputs from the NULP such as textbooks (Glewwe, Kremer, and Moulin 2009) and teacher training (Cilliers et al. 2019).[15] These packaged interventions show real promise, with RCTs sometimes finding effects as large as those found for the NULP (e.g., Gove et al. (2017)). The PRIMR intervention has larger effects when implemented in the students' mother tongue, but only for literacy in that language (Piper et al. 2018). Delavallade, Griffith, and Thornton (2019) point out that the majority of programs actually *implemented* in developing countries involve a packaged bundle of education inputs.

Despite extensive research evaluating the effectiveness of educational inputs, the existing literature has had little to say about treatment effect heterogeneity. Most studies concentrate on presenting average treatment effects on test scores. Among studies that examine individual components of the NULP, Jackson and Makarin (2018) use a conditional quantile treatment regression approach to show that the lesson plans matter more for weaker teachers, while Glewwe, Kremer, and Moulin (2009) find that textbooks only improve scores for the strongest students.

3 Evaluation

3.1 Evaluation design

The evaluation of the NULP took place over four academic years running from 2013 through 2016. The evaluation involved 38 schools in 2013 with an additional 90 schools added in 2014. The evaluation assigned eligible government primary schools at random to one of three treatment arms: the full-cost NULP treatment described in the preceding section, a reduced-cost version of the NULP treatment designed to approximate what a scaled-up, less expensive, government-operated version of the program would look like, and a business-as-usual control condition.[16] Randomization took place within pre-defined groups (strata) of

[15] There has been limited research evaluating the components of the NULP in developed-country contexts. An exception is Jackson and Makarin (2018), which studies the effect of one of the NULP's components, providing pre-designed lesson plans in schools in Virginia. Giving teachers access to pre-designed lesson plans has small but statistically significant effects on math test scores, effects that increase with additional support to use the lesson plans. There is also an extensive literature on the effects of mother-tongue instruction in the US, focused primarily on Spanish-language immersion courses (see, e.g., Rossell and Baker (1996)).

[16] School eligibility differed slightly in the first two years of the study. In 2013, eligibility required that a school have two P1 classrooms, lockable classrooms, a head teacher regarded as "engaged", less than 135 students/teacher, and be located less than 20km from the main government coordinating center. In 2014 the study team dropped most of the requirements, demanding only that schools have less than 150

three schools.[17] The reduced-cost version embodied two main changes: 1) instead of Mango Tree staff directly providing teacher training and teacher support, Ministry of Education coordinating center tutors provided it via a "cascade" or "training-of-trainers" model; and 2) teachers received fewer support visits throughout the year.[18] In other words, the three arms vary the intensity of the treatment across schools in a way that varies across dimensions of the package treatment.[19]

The NULP program was provided to P1 teachers in treatment schools in 2013 and 2014. In 2015 the program was then provided to P2 teachers in treatment schools, and in 2016, the program was provided to P3 teachers in treatment schools

3.2 Analytical Sample

In this paper, we focus solely on students who entered first grade in 2014 in one of the 128 study schools. Because the intervention was rolled out to grades P1, P2 and P3 across years, these students were exposed to three full academic years of whatever treatment their school received. By focusing on just one cohort of students, we avoid mechanical variation in treatment intensity resulting from differing amounts of exposure. Buhl-Wiggers et al. (2018a) study the effects of the varying exposure to the program.

Funding limitations prevented doing full data collection on all of the students who started P1 in the study schools in 2014. Instead, the evaluation sampled 100 P1 students from each school; in schools with fewer than 100 P1 all available P1 students were included in the evaluation. Students were sampled from classrooms in two ways. First, at the beginning of the school year, we drew an initial sample, of 40 P1 students from each of the original 38 schools and 80 P1 students from each of the additional 90 schools. Second, at the end of P1, we drew a top-up sample of 60 students at the original 38 schools and of 20 students from the additional 90 schools. Both samples were stratified by sex and classroom.[20]

To reach what we call our "main analysis sample" we impose two additional restrictions.

students/teacher and be located at most 22km from a government coordinating center.

[17] Groups are defined based on P1 enrollment, coordinating center, and distance to coordinating center headquarters.

[18] Buhl-Wiggers et al. (2018a) describe the lessons the NULP evaluation provides regarding program scale-up.

[19] Kerwin and Thornton (2020) discuss implementation and implementation fidelity in some detail.

[20] The top-up sample differs somewhat from the initial sample on exogenous characteristics, reflecting the fact that they are sampled from students who enrolled in school later. They are about half a year older and 1.7 percentage points more likely to be female. Our top-up sample has the disadvantage that it could reflect systematic differences in attendance at the end of the school year driven by the treatment arm assigned to the school. We looked for this in the data but found no evidence of such systematic differences (not shown). Moreover, we obtain qualitatively similar results using just the initial sample and just the top-up sample (available by request).

First, we require the students to have valid test scores at the end of P3 (in 2016) for use in constructing the outcomes we study. Second, because we devote a good portion of our attention to the analysis of moderators, we require complete data on all but one of the variables we use as moderators.[21] The one exception is baseline test scores (measured in P1, at the beginning of 2014); we do not condition on the presence of this variable to avoid losing all of the top-up sample observations, which lack baseline test scores because they were sampled at the end of the 2014 school year. In the analyses that includes student baseline test scores, we recode missing values to zero and include an indicator for missing values.[22]

Imposing these restrictions yields a main analysis sample of 4,868 students, with 1,427 in the control group, 1,681 in the full-cost treatment group and 1,760 in the reduced-cost treatment group. Online Appendix Table 1 provides further details on the construction of the main analysis sample, including the number of observations lost due to each restriction we impose.[23]

3.3 Learning Outcomes

Our outcome measure captures student performance in reading Leblango, the local language of students in our study schools. In particular, we construct an index built on scores from the Early Grade Reading Assessment (EGRA) administered to students at the end of P3.[24] The EGRA is an internationally standardized exam—externally validated in Leblango—intended to evaluate reading skills (RTI International 2009). The exam consists of six components: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. Following earlier research on the NULP by Kerwin and Thornton (2020), we construct a principal component score index for the entire exam using the factor loadings from the control group. We use the combined score, standardized with respect to the control group in P3, as our primary outcome variable throughout.[25]

[21] We measure all moderators at the beginning of the study in 2014, with the exception of the teacher characteristics, which are for the P3 teachers in 2016. We choose a single set of teachers to reduce the dimensionality of the data, and focus on the most-recent teachers as they are the most proximal influences on the endline test scores (and also because their data has the fewest missing values).

[22] Almost all students (87% of the sample) received a zero on their 2014 baseline test score (not shown).

[23] We lose somewhat more observations from the control group than the other two study arms due to missing moderators.

[24] Only students who were present in school were administered the endline exam. See Online Appendix Table 1 for a breakdown of the number of observations we lose due to missing endline scores.

[25] We repeated many of our analyses using EGRA English scores as well as math scores. These alternative outcomes yield similar qualitative conclusions. Note that impacts on test scores at the end of P3 capture the effects of exposure to treatment for three years, in P1, P2, and P3.

3.4 Covariates

We divide our covariates into three sets based on their level of variation: student characteristics, teacher characteristics, and school characteristics. For students, our covariates comprise baseline test scores (setting missing values to zero) and an indicator for a missing baseline score, a male indicator, and student age at baseline.[26] For teachers, they include a male indicator, age, years of teaching experience, and their years of completed schooling. For schools, we have total enrollment from P1-P7, the total number of teachers from P1-P7, the overall P1-P3 pupil-to-teacher ratio, and the pass rate on the Primary Leaving Exam (PLE) in the year before the intervention started.[27]

Online Appendix Table 2 presents covariate means by study arm along with balance tests. We see that 50 percent of the students in the sample are girls—by construction, due to our sampling strategy—and the average student age is between eight and nine years old. Teachers average 39 to 43 years of age, 15 years of experience, and 14 years of education. Schools average around 900 total students, 14 teachers, a P1-P3 pupil-teacher ratio around 67, and a PLE pass rate of a little less than 5 percent. Students in the two treatment arms are more likely to have male teachers and their teachers are less experienced and slightly younger; students in the reduced-cost study arm attend schools that are somewhat larger overall. Following Deaton and Cartwright (2018), who gently (and rightly) mock the epistemology implicit in taking significance tests of nulls known to hold in the population too seriously, we focus mainly on the magnitudes of the sample covariate imbalances, which are small and unremarkable for most variables. Students in the two treatment arms are more likely to have male teachers and their teachers are less experienced and slightly younger; students in the reduced-cost study arm attend schools that are somewhat larger overall.

4 Average Treatment Effects

As a point of comparison, we begin by estimating the average effect of each version of the NULP in our main analysis sample. We estimate the following linear model:

$$Y_{isc} = \beta_{FC} FC_s + \beta_{RC} RC_s + \beta_Y Y_{i,P1} + \beta_D D_i + \alpha_c + \epsilon_{isc} \tag{1}$$

In equation (1), Y_{isc} denotes the outcome (i.e. the reading in Leblango test score index just defined) for student "i" in school "s" in stratification cell "c". FC_s and RC_s indicate

[26] We address outliers (which we suspect include both measurement error and real exceptional cases) by censoring the data at 7 and 13; this affects just 0.39% of all observations.

[27] Students are perfectly nested within schools, but each student was exposed to up to three different teachers during the program; we use the characteristics of their P3 teachers.

assignment to the full-cost and reduced-cost treatment arms for school "s", respectively, which implies that β_{FC} and β_{RC} represent the average treatment effects of the two versions. $Y_{i,P1}$ denotes the baseline (i.e. P1) Leblango reading test score index while D_i is an indicator equal to one when the baseline score is missing (and so set to zero, as noted above). Finally, α_c is a treatment stratification cell fixed effect and, as always, ϵ_{isc} is a mean-zero term that captures the effects of all omitted determinants of test scores. We cluster the standard errors at the school level given the school-level treatment assignment.

Table 1 presents the estimates from four versions of equation (1). Column (1) shows unconditional treatment effects (i.e. the simple mean difference). Column (2) adds the stratification cell fixed effects a_c. As expected based on, e.g. Bruhn and McKenzie (2009), adding these fixed effects improves the statistical efficiency of our estimates. More importantly, we require them for consistency because we have different shares of schools and students in each study arm in each stratification cell. Column (3) adds controls for students' baseline test scores, dropping students with a missing test score, while column (4) keeps the students with missing baseline test scores and includes an indicator for missing values.[28]

[Table 1 about here.]

We estimate average treatment effects of 1.40 SDs for the full-cost treatment and 0.74 SDs for the reduced-cost treatment using our preferred specification in column (4).[29] The results vary only slightly across columns, ranging from 1.40 to 1.53 for the full-cost version and from 0.74 to 0.80 for the reduced-cost version. This robustness motivates our choice to use the specification in column (4) as our main specification for the remainder of the paper. Substantively, these represent very large impacts. The full-cost program effect sits in the 99$^{\text{th}}$ percentile of the overall distribution of impacts of the primary-school education programs reviewed in the McEwan (2015) meta-analysis. Moreover, not a single program in his study had such a large effect on reading scores. Even the reduced-cost program effects are large relative to the literature; for example, 95% of the experiments in McEwan yield treatment effects below 0.45 SDs, with the average being 0.10 SDs.

[28] Fans of Freedman (2008) will prefer Column (2) while fans of Lin (2013) will prefer column (4). We tend to agree with the latter but offer both sets of estimates in this table in the spirit of celebrating our (epistemological) diversity.

[29] Buhl-Wiggers et al. (2018a) report an average treatment effect for the full-cost version of 1.35 SDs. The difference springs from slightly different samples in the two cases, with their sample requiring non-missing data on a smaller set of conditioning variables.

5 Establishing treatment effect heterogeneity

This section pursues the first stage of our analysis by interrogating the data for evidence of treatment effect heterogeneity using the classical statistical bounds that rely only on the information in the marginal outcome distributions.

5.1 Formalities

The FH bounds capture the limits on $F(Y_1, Y_0)$, the joint CDF of the outcome under the treated state, Y_1, and the control state, Y_0, implied by their marginal distributions. Put differently, the FH bounds define the set of identified joint distributions consistent with given marginal distributions without the addition of any further identifying information. In the context of our three-armed experiment, the treated state could represent either the full-cost version of NULP or the reduced-cost version of NULP.

For continuous variables, the Fréchet-Höffding bounds are:

$$\begin{aligned} & max[F_1(Y_1|D=1) + F_0(Y_0|D=1) - 1, 0] \\ & \leq F(Y_1, Y_0|D=1) \\ & \leq min[F_1(Y_1|D=1, F_0(Y_0|D=1)] \end{aligned} \tag{2}$$

where $F_1(.)$ is the marginal distribution of the outcome variable in the treated state and $F_0(.)$ is the marginal distribution in the control state. The lower bound corresponds to the case of perfect negative dependence or "rank inversion" as it implies a rank correlation of -1.0. The upper bound corresponds to perfect positive dependence or "rank preservation" as it implies a rank correlation of 1.0. Thinking in terms of ranks helps illustrate the intuition that underlies the FH bounds. In rank preservation, the CDF implicitly links a given rank in one outcome distribution with the same rank in the other outcome distribution, so that, for example, the counterfactual for a student at the 90^{th} percentile of the full-cost program outcome distribution equals the 90^{th} percentile of the control outcome distribution. In contrast, with rank inversion the counterfactual for a student at the 90^{th} percentile of the full-cost program outcome distribution equals the 10^{th} percentile of the control group outcome distribution.

Cambanis, Simons, and Stout (1976) show that all super-additive and sub-additive parameters obtain their extreme values at the FH bounding distributions. Tchen (1980) shows that Spearman's ρ and Kendall's τ do too. The class of super-additive parameters includes the Pearson correlation, which, as HSC (1997) point out, implies that the FH bounding distributions also bound the treatment effect variance, $var(Y_1 - Y_0)$. To see the intuition,

suppose that $F_1 \sim U[0,1]$ and $F_0 \sim U[0,1]$, i.e. both outcomes have uniform distributions on the unit interval. The FH upper bound distribution, and its attendant rank preservation, then has $Y_1 = Y_0$ so that the variance of the treatment effects equals exactly zero in the population. In contrast, at the FH lower bound distribution, with its attendant rank inversion, treatment effects decrease linearly from 1.0 to -1.0 as Y_1 moves from 1.0 to 0.0, so that the treatment effect variance well exceeds zero (and, indeed, obtains its maximum consistent with the given uniform marginals).[30]

5.2 Implementation

We start by collapsing the outcome distributions for the three treatment arms into percentiles, both to simplify the computations and because the three arms contain different numbers of students. The FH upper bound distributions, which embody rank preservation, then match percentiles between one of the treatment arms and the control arms. Subtracting the control percentile outcome from the treated percentile outcome gives the treatment effect for that percentile for that version of the program. A similar operation, but with the control outcome percentiles inverted, provides the treatment effects associated FH lower bound distribution, which embodies rank inversion.[31]

For each combination of bound (upper or lower) and treatment arm (full-cost or reduced-cost) we calculate the Pearson correlation between the percentiles of the treated and control outcome distributions. In addition, we calculate the impact standard deviation as the square root of the variance of the percentile-specific impacts and the fraction with a positive impact as the fraction of non-negative percentile-specific impact estimates.[32] We compute standard errors using the non-parametric bootstrap, drawing 1000 samples of students—each with as many observations as the original sample—with replacement from the main analysis sample and then repeating the entire exercise just described for each bootstrap sample.[33] The standard deviation of the 1000 bootstrap estimates provides our bootstrap standard errors.

[30] While HSC (1997) are correct when they state that "[t]these inequalities [the FH bounds] are not helpful in bounding the distribution of [the treatment effects]" the marginal distributions do provide some information about this distribution: see, e.g. Williamson and Downs (1990) and Fan and Park (2010).

[31] One could imagine related exercises such as imposing the bounds within stratification cells or imposing them after subtracting off stratification cell fixed effects from all of the outcomes.

[32] Calculating the impact variance using the percentiles, rather than some finer approximation to the outcome distributions, likely leads to a mild understatement of the true population bounds on the impact variance.

[33] Note that we do not sample schools and then students within schools in our bootstrap for computational simplicity. As a result, we likely somewhat understate the sampling variability in our estimates.

5.3 Findings

Table 2 presents estimates of the various statistics associated with the FH bounding distributions. Columns (1) and (2) relate to the full-cost treatment and Columns (3) and (4) relate to the reduced-cost treatment. For each treatment, the left-hand column gives statistics under rank preservation (the FH upper bound distribution) and the right-hand column gives statistics under rank inversion (the FH lower bound distribution). Each column provides the treatment effects associated with the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles of the control-group outcome distribution, along with the fraction of positive percentile impacts, the impact standard deviation, and the outcome correlation.

We focus here on the bottom three rows of the table and defer discussion of treatment effects at particular quantiles of the outcome distribution to Section 6.2, which considers such quantile treatment effects in detail. We begin with the Pearson correlation in the last row where super-additivity implies that our estimates represent bounds. By construction, rank preservation yields a large positive outcome correlation in both cases, while rank inversion yields a large negative one. At the same time, while the rank correlations equal 1.0 and -1.0, the Pearson correlations do not, reflecting the interaction of the different formulae underlying the two correlation measures and the shapes of the outcome distributions.

As noted above, super-additivity also implies that we estimate bounds on the impact standard deviation, where the lower bound is obtained under rank preservation and the upper bound under rank inversion. We find bounds of (1.066, 2.615) for the full-cost program and of (0.642, 2.219) for the reduced-cost program. These lower bounds are huge! The lower bound for the full-cost program equals 76 percent of its mean impact in Column (4) of Table 1; similarly, the lower bound for the reduced-cost program equals about 87 percent of its mean impact. If we assume normally distributed impacts this implies that for the full-cost program 29 percent of the students have impacts of at least 2.0 SDs while only 10 percent have negative impacts.

Another way of looking at our lower bound estimate of the impact standard deviation for the full-cost program compares it to cross-program variation in average treatment effects. Staying within our data, and again assuming normality, the difference between the 5^{th} and 95^{th} percentile treatment effects equals 3.5 standard deviations, and thus is over four times the difference in average impacts between the full- and reduced-cost programs. It also far exceeds the difference in average treatment effects between the most- and least-effective programs among the 76 randomized experiments covered in the McEwan (2015) systematic review of primary education interventions in the developing world. Those interventions vary in their mean impacts from -0.57 to 1.51 SDs, a range of 2.08 SDs.

Finally, consider the fraction positive. It is not super-additive, and so need not fall into

the range defined by our bounds. At the same time, our bounds on the fraction positive further illustrate the underlying intuition of the FH bounds. Consider the example given above wherein both Y_1 and Y_0 have $U[0, 1]$ distributions. In this case, rank preservation yields a fraction positive (really non-negative as we define it) of 1.0 because $Y_1 = Y_0$ so that all of the treatment effects equal zero. In contrast, rank inversion yields a fraction positive of 0.5, as the bottom half of the treated units get linked to the top half of the untreated units and vice versa. More generally, rank inversion necessarily leads to at least some fraction of the treated units having negative treatment effects so long as the two distributions share some common support. To see this, first change the example so that the treated unit outcomes are distributed $U[0.9, 1.9]$. This yields a fraction positive of 0.95 under rank inversion, as only those treated units with outcomes in [0.90, 0.95] get linked to control outcomes that exceed their own. Changing the example again so that the treated outcomes are distributed $U[1.1, 2.1]$ implies a fraction positive of 1.0 even under rank inversion, because every treated outcome with positive support exceeds every control outcome with positive support.

In our data, with the treated outcomes well above the control outcomes on average for both versions of the NULP program, we find that nearly 100 percent of students experience positive treatment effects in both treatment arms under rank preservation; even under rank inversion, that fraction only falls to almost 0.70 for the full-cost program and 0.65 for the reduced-cost program.

[Table 2 about here.]

5.4 Testing the null of a common treatment effect

In the preceding section, we carefully avoided using the bootstrap standard errors to perform a simple test of the null of a zero impact standard deviation based on the ratio of one of the estimated impact standard deviation lower bounds to its bootstrap standard error. We did so because HSC (1997, Appendix E) makes a strong case that the bootstrap standard errors, though they do a reasonable job when the population impact standard deviation differs non-trivially from zero, do a very poor job when it equals zero, its value under the common effect null. Online Appendix Table 3 repeats a subset of that analysis using our data; our results imply the same qualitative conclusion.

The statistics literature talks about the general problems that arise when testing nulls that lie at the boundary of the parameter space. In our context, variances must, by construction, lie in the interval $[0, \infty)$. Our null of zero lies on the edge of that set. To see the problem at a very prosaic level, think about a sample from an RCT from a population

and intervention where the null holds for some outcome. Imagine calculating the impact standard deviation using said sample as we do above. Due to sampling variation, the impact standard deviation will exceed zero with probability one, because with probability one at least one of the percentile differences will not equal zero due to sampling variation.

We address both the boundary issue and the issue with the bootstrap standard errors by using the randomization inference procedure developed in Appendix E of HSC (1997).[34] Intuitively, their test constructs an estimate of the sampling distribution under the null via resampling from the experimental control group. Because no control group members receive the treatment, the null holds in resamples from the control group wherein we construct impacts via randomly assigned faux treatment and control groups. By repeatedly drawing samples of the same size as our original data, dividing them at random into faux treatment and control groups, and estimating the impact standard deviation as we did using the original data, we can approximate the sampling distribution of the estimated impact standard deviation under the null. Our test then locates the estimated lower bound impact standard deviation from the actual data within the estimated sampling distribution. If it, for example, lies above the 95^{th} percentile of that distribution, we can reject the null at the 0.05 level, and so on.

We implement the test as follows: First, we sample with replacement from the control group in the main analysis sample. Then we randomly sort the sample and assign the first half of the observations to the faux control group and the second half to the faux treatment group.[35] We add the average treatment effect from Table 1 to the test scores for the faux treatment group.[36] Using this synthetic dataset, we collapse the treatment and control outcome distributions into percentiles. We take the differences across percentiles and compute the standard deviation of these differences, exactly as we did using the original analysis sample.

We repeat this process 10,000 times, once for a sample of size of 3,108 with 1,427 controls, which corresponds to the full-cost versus control contrast, and once for a sample size of 3,187 with 1,427 controls, which corresponds to the reduced-cost versus control contrast. This yields two empirical distributions of 10,000 estimated impact standard deviations generated under the null hypothesis of zero variance of impacts. Online Appendix Table 4 presents cutoff values under the null. Comparing the impact standard deviations we obtain using the

[34] Note that HSC (1997) do not use the term "randomization inference" to describe what they do, as that term had not yet entered general circulation in economics.

[35] When drawing the bootstrap samples, we do not cluster by school, i.e. we resample individual students and not entire schools and then students within schools. Implicitly, this means that our null holds the set of schools fixed.

[36] Adding the average treatment effect for either program version (or indeed any other constant) to the faux treatment group outcomes does not change our findings because it does not change the impact variance.

original data to the cutoff values corresponding to a p-value of 0.0001 shows that we can easily reject the null at that level for both program versions. Thus, our data provide substantially stronger statistical evidence against the common effect null than the data employed in either HSC (1997) or Djebbari and Smith (2008).

6 Exploring treatment effect heterogeneity

This section pursues the second stage of our analysis by examining the extent to which additional assumptions, first stochastic increasingness and then rank preservation, reduce and clarify the variation revealed by the classical bounds.

6.1 FL bounds

6.1.1 Introduction

The FH bounds tell us that our data embody a great deal of treatment effect heterogeneity with, for example, bounds on the standard deviation of $(Y_1 - Y_0)$ of 1.07 and 2.62 for the full-cost NULP program. In this section, we consider the alternative bounds developed in FL (2020). They show that limiting consideration to joint distributions of potential outcomes that exhibit the property of "Mutually Stochastic Increasingness" (MSI) allows for informative pointwise bounds on the average treatment effects at specific quantiles of the potential outcomes.[37]

In their words, MSI implies that "the distribution of outcomes under treatment among individuals who would have realized a higher outcome in the control state, (weakly) stochastically dominates the distribution among individuals who would have realized a lower outcome in the control state, and vice versa." In our setting, this means that if student A has a higher test score than student B under the status quo, student A will also probably have a higher score than student B in the treated state of the world, and similarly for student B if their roles are reversed. MSI implies a positive rank correlation, which links our analysis to that in Tables 5A and 5B in HSC (1997), which presents descriptive statistics on distributions of impacts randomly sampled conditional on particular values of the rank correlation between Y_1 and Y_0.[38] While MSI does not imply a specific value of the rank correlation, it does rule

[37] In addition to the pointwise conditional bounds, Section 3.1.2 of FL (2020) also offers two sets of bounds on the overall treatment effect distribution. We do not investigate them here mainly because we have failed to cajole their software into producing reasonable estimates on our data.

[38] In contrast, a positive rank correlation does not imply MSI. To see this, suppose again that Y_1 and Y_0 have $U[0.0, 1.0]$ marginal distributions. Now imagine that the joint distribution has $Y_1 = Y_0 + 0.1$ for $Y_0 \in [0.0, 0.9]$ and $Y_1 = Y_0 - 0.9$ for $Y_0 \in [0.9, 1.0]$. This joint distribution clearly has positive rank correlation as the ranks move in lockstep for 90 percent of the population, but not MSI because for the units at the top

out all negative values. MSI differs from rank preservation in that the latter implies a rank correlation of one—the best student under the control state of affairs is also the best student when the treatment is applied, and likewise for every rank—while the former allows any positive rank correlation.

Does MSI make sense in our substantive context? HSC (1997) point out that MSI follows naturally when participants have some knowledge of their potential outcomes and self-select into an intervention. As we study (essentially) mandatory programs, we cannot use this argument to justify MSI in our context. FL (2020) argue in their context—charter schools in the U.S.—that many drivers of academic performance—such as students' latent ability and effort—likely imply better performance in both the treated and untreated states. In a loose sense, their many drivers represent micro-foundations for a one factor ("ability") model with noise, a model we find quite plausible in our context. At the same time, one worry is that our data may contain some students who flourish in the control world of call-and-response in English and flail in the NULP world of scripts and slates and clocks in Leblango, or the reverse. Too many such students would imply that MSI fails even as an approximation.

6.1.2 Formalities and implementation

FL (2020) define the potential outcomes Y_1 and Y_0 as mutually stochastically increasing if the following property holds:

$Pr(Y_1 \leq t|Y_0 = y)$ and $Pr(Y_0 \leq t|Y_1 = y)$ are each non-increasing in y almost everywhere.

In words, this means that if one student has a higher outcome in the control state of the world, her conditional distribution of outcomes in the treated state first-order stochastically dominates that of a student with a lower outcome in the control state. Under this assumption, they show that the lower-bound CDF is given by

$$F^L_{\Delta|Y_0}(t|Y_0) = \begin{cases} 0, Y_0 > F_0^{-1}(F_1(Y_0 + t)) \\ \dfrac{F_1(Y_0 + t) - F_0(Y_0)}{1 - F_0(Y_0)}, Y_0 > F_0^{-1}(F_1(Y_0 + t)) \end{cases} \tag{3}$$

And the upper bound is given by

$$F^U_{\Delta|Y_0}(t|Y_0) = \begin{cases} \dfrac{F_1(Y_0 + t)}{F_0(Y_0)}, Y_0 > F_0^{-1}(F_1(Y_0 + t)) \\ 1, Y_0 > F_0^{-1}(F_1(Y_0 + t)) \end{cases} \tag{4}$$

of the untreated outcome distribution, things only get worse with treatment.

These expressions give the probability that the treatment effect is less than or equal to a given value, t. To compute the bounds, we need to estimate the unconditional CDFs, $F_0(.)$ and $F_1(.)$. The FL (2020) algorithm for this proceeds as follows: First, compute $F_0(y+t)$ as the sample mean of the indicator $1(Y_i \leq y+t)$ in the control group data. Similarly, compute $F_1(y+t)$ as the sample mean of the indicator $1(Y_i \leq y+t)$ in the treatment group data. Then plug those estimates into equations (3) and (4) to compute estimates of the lower- and upper-bound conditional CDFs. Finally, use these estimated CDFs to compute lower and upper bounds on the conditional (i.e. quantile-specific) treatment effects:

$$\Delta^L(Y_d) = \int t dF^U_{\Delta|Y_d}(t|Y_d), \tag{5}$$

$$\Delta^U(Y_d) = \int t dF^L_{\Delta|Y_d}(t|Y_d) \tag{6}$$

Intuitively (though not obviously), the pointwise lower bound on the conditional treatment effect in (5) corresponds to a joint distribution with rank preservation above the evaluation point and independence of the treated and untreated outcomes below the evaluation point. Similarly, the pointwise upper bound on the conditional treatment effect in (6) has rank preservation below the evaluation point and independence above it. Practically, computing the underlying conditional CDFs requires numerical integration of the numerical derivatives in (5) and (6), so it proceeds slowly in our data, which has many more observations than the application in FL (2020).[39]

6.1.3 Findings

We present the pointwise FL bounds on the conditional expected impacts in Table 3. Columns (1) and (2) present the lower and upper bounds for the effects of the full-cost program by control-group percentile, while Columns (3) and (4) present the bounds for the reduced-cost program. The table shows that the mean effects of the full-cost program could range from 0.20 SDs to 2.65 SDs for the 5$^{\text{th}}$-percentile student, and from -0.57 SDs to 4.27 SDs for the 95$^{\text{th}}$ percentile student. For the reduced-cost program, the 5$^{\text{th}}$-percentile student on average gains between 0.16 and 1.92 SDs, and the 95$^{\text{th}}$-percentile student sees mean effects that range from a 1.19-SD *loss* to a 3.23-SD gain. The upper bounds increase monotonically

[39] The FL bounds take approximately 10 days to run for each treatment arm on our sample, using an Intel Core i7-4790 3.6 GHz CPU with eight cores and 32 GB of RAM, Stata 16.0, and the replication code from their paper. We can estimate the bounds for the two treatment arms simultaneously using separate instances of Stata, but the code is not parallelized, so each instance of the program uses just one core. Drawing samples from simulated distributions that impose a positive rank correlation as in HSC (1997) might speed things up.

with percentiles of the control-group outcome distribution for both program variants, while the lower bounds initially rise and then fall for the highest percentiles.

We highlight three features of the FL bounds: first, all the bounds turn out quite wide in a substantive sense. For example, for the full-cost program, the average treatment effect for the student at the median of the control distribution has a range of over three SDs. Second, unlike the FH bounds, the pointwise FL bounds do not allow us to rule out the common effect model (or even its expected value analogue) as a wide range of expected treatment effects lie within all of the pointwise bounds. Third, and perhaps most useful, the FL bounds tell us that only in the very upper percentiles of the control state outcome distribution do students have any possibility of negative average treatment effects for either the full-cost or the reduced-cost version of the NULP. MSI does have some valuable substantive bite.

[Table 3 about here.]

6.2 Quantile treatment effects

6.2.1 Introduction

We now impose an even stronger assumption than stochastic increasingness, namely rank preservation. As described in Section 5.1, the FH upper bound distribution implicitly embodies rank preservation, so that the rank correlation between treated and control outcomes equals one in the population. An alternative conceptual and computational path to the FH upper bound distribution leads through the estimation of Quantile Treatment Effects (QTEs).[40] In the context of an experiment (so that we need not worry about selection into treatment and its attendant biases) the quantile treatment effects consist of the simple differences in quantiles between the treatment group outcome distribution and the control group outcome distribution.

These QTEs admit of two distinct interpretations. The first interpretation does not impose rank preservation but instead remains agnostic about the underlying joint outcome distribution. Under this interpretation, the QTEs inform the researcher about the effect of treatment on the shape of the outcome distribution and related parameters. For example, a pattern of negative QTEs at low quantiles and positive QTEs at high quantiles implies that

[40] Koenker and Bassett (1978) began the literature on quantile regression in economics. Important early applications in program evaluation contexts include Lehmann and D'Abrera (1975) and Doksum (1974) in the statistics literature and HSC (1997), Koenker and Bilias (2002), Abadie, Angrist, and Imbens (2002) and Bitler, Gelbach, and Hoynes (2006) in the economics literature. HSC (1997) do not use the term QTE because it has not yet entered the applied econometric lexicon when they wrote.

the treatment increases the outcome variance. Graphing the QTEs against the percentiles can add meaningfully to the information provided by the average treatment effect. Indeed, it surprises us that such graphs have not become routine in experimental evaluations.

The second interpretation presumes rank preservation and so returns us to the world of the FH upper bound distribution. In this interpretation, the QTEs represent impacts *at quantiles* as well as impacts *on quantiles* as in the first interpretation. Put differently, under rank preservation the QTE for the, say, 75$^{\text{th}}$ quantile indicates the impact on students at the 75$^{\text{th}}$ quantile. Thus, we can make statements such as "the treatment improves the test score of the X$^{\text{th}}$ percentile student by Y SDs." The first interpretation does not allow such statements, because under the first interpretation the joint distribution could be anything (consistent with the given marginals).[41]

6.2.2 Implementation

We estimate QTEs for each quintile from the 5$^{\text{th}}$ to the 95$^{\text{th}}$ percentile using the estimator defined in Koenker and Bassett (1978) as embodied in Stata's `qreg` command.[42] We present bootstrap standard errors based on 250 replications, clustered by school and stratified by stratification cell (i.e. resampling schools from within their original stratification cells rather than drawing them from the entire original sample). Our figures present the quantile regression point estimates as a connected black line, with 95% confidence intervals in gray. For reference, we also show the average treatment effects from Column (1) of Table 1 on the figures; these average effects correspond most closely to our QTEs, which also do not control for the stratification cell indicators.

We take advantage of the QTE framework (and of Stata's `sqreg` command) to conduct an alternative test of the common treatment effect null. More precisely, we test an implication of that null: namely, the equality of the QTEs at various percentiles. This null is implied by, but does not imply, the null of the common effect model, as one can imagine forms of treatment effect heterogeneity consistent with equal QTEs. Thus, rejections using this test statistically imply a non-zero impact variance but failure to reject does not imply an impact variance of zero. Our test focuses on QTEs at the 5$^{\text{th}}$, 10$^{\text{th}}$, 15$^{\text{th}}$,..., 95$^{\text{th}}$ percentiles.

[41] Bitler, Hoynes, and Domina (2016) compare the knowledge produced by quantile treatment effects and by subgroup impacts with subgroups defined based on baseline outcomes.

[42] HSC (1997), who did not make the connection to quantile regression, construct their QTEs via percentile differences, calculating standard errors using the method in Csörgo (1983).

6.2.3 Findings

Figure 1 shows the quantile treatment effect estimates. Both program variants exhibit monotonically increasing treatment effects across the quantiles of the outcome distribution. We see no effect of the two program versions on the 5^{th} percentile of outcomes.[43] However, the QTEs increase steadily up to about 2.97 SDs on the 95^{th} percentile for the full-cost version and about 1.87 SDs for the reduced-cost version. Even without rank preservation, this pattern implies that both versions of NULP strongly increase the variance and inequality of academic outcomes as well as the mean. Adding rank preservation adds the further implication that the treatment effect strongly increases in student test performance in the control state. Put differently, students who would struggle under the existing regime would also struggle under both variants of NULP, while students who do (relatively) well under the current regime as embodied in the control state would do much better under NULP.

Brief visual consideration of Figure 1 makes it clear how our formal test of the null of equal quantile treatment effects will turn out. As confirmed by the exact test statistics presented in Online Appendix Table 5, we easily reject that null even at the 0.001 level for both the full-cost and reduced-cost programs.

[Figure 1 about here.]

6.2.4 Testing rank preservation

Because we cannot ever know the joint distribution of Y_1 and Y_0, the assumption of rank preservation is fundamentally untestable. But, helpfully, Bitler, Gelbach, and Hoynes (2005) point out that it does have testable implications.[44] Under rank preservation, characteristics of units not affected by treatment should look the same at corresponding quantiles of the treatment and control outcome distributions. For example, under rank preservation, the demographic characteristics of students at the 75^{th} percentile of the control outcome distribution should mirror those of students at 75^{th} percentile of the reduced-cost program outcome distribution and of students at the 75^{th} percentile of the full-cost program outcome distribution. As with our test of equal QTEs, because we test an implication of the null of

[43] The low impacts at the 5^{th} percentile result in part from the fact that nearly 10 percent of the control group scores zero on the entire Leblango EGRA while the 5^{th} percentile scores in the two treatment arms differ only marginally from zero. In one sense, this is a "floor" effect but in another sense it is not as the test provides a clear indication that these students have learned very little about how to read Leblango after three years. Note that the early parts of the test are very easy—getting a score of zero requires that students are unable to recognize even a single letter of the alphabet.

[44] We cite the working paper version of this paper because some misguided editor demanded that the authors drop the test from the published version.

interest rather than the null itself, rejection of the null of characteristic balance by outcome quantile allows us to infer that rank preservation does not hold, but failure to reject does not allow us to infer that it does hold. Of course, magnitudes matter as well as test statistics. A mild statistical rejection of balance combined with relatively small substantive differences could support an interpretation that rank preservation holds in some approximate sense (e.g. with a rank correlation around 0.9).

Our implementation generally follows Djebbari and Smith (2008) who in turn followed the original scheme in Bitler, Gelbach, and Hoynes (2005). First, we divide the outcome (i.e. our 2016 endline Leblango reading score index) into quartiles separately by treatment arm. The choice of quartiles, rather than, say, quintiles or deciles or halves, embodies a tradeoff between fidelity to the null and the power of the test. Strictly speaking, the null concerns covariate balance at specific quantiles of the outcome distribution. The test concerns balance within intervals of quantiles because a test at a specific quantile would have no power. Increasing the width of the test interval increases statistical power while at the same time reducing the correspondence between the null implicit in the test and the null of covariate balance at specific quantiles. Within each quartile, we regress 12 different covariates (baseline test score, an indicator of missing baseline test score, student gender, student age, teacher gender, teacher age, teacher experience, teacher education, school enrollment, pupil-teacher ratio, PLE pass rate, number of teachers) on indicators for the two treatments, controlling for stratification cell fixed effects.[45] These coefficients on the treatment indicators represent the quartile-specific mean differences in the covariate. Under rank preservation, they should equal zero up to sampling variation.

We construct our bootstrap confidence intervals for the null of zero differences in two steps. In the first step, we draw a bootstrap sample of schools with replacement from each treatment arm. We combine these schools and then randomly reassign them to create a faux control sample, a faux reduced-cost sample, and a faux full-cost sample, with the same proportion of schools in each as in the original data. In the second step, we sample students with replacement from the bootstrap sample schools. Using the resulting bootstrap sample of students, we repeat the covariate balance regressions we performed on the original data and save the estimates. We repeat this exercise 1,000 times, sort the resulting estimates, and then use, for example, the 25^{th} and 975^{th} largest estimates as the bounds for the 95 percent confidence interval.

Table 4 presents the results from this exercise. We easily reject rank preservation. At the

[45] We modify the procedure in Bitler, Gelbach, and Hoynes (2005) by adding a step in which we subtract off the overall average effect of each treatment (across all four quartiles) on the characteristic of interest. This focuses the test on changes in ranks rather by removing the small amounts of imbalance that result from sampling variation.

10 percent level, we reject the null in 14 out of 48 tests for the full-cost program, and 13 out of 48 tests for the reduced-cost version. We would expect a total of about five rejections for independent tests; our tests are not independent which implies that we should expect even fewer rejections.[46]

A natural model that implies rank preservation assumes that test scores result from a single underlying factor—call it "ability"—with observed scores in each treatment arm a strictly increasing function of ability. Adding a bit of measurement error in the tests implies that rank preservation holds only approximately, with the strength of the approximation depending on the signal-to-noise ratio of the test. We can shed some additional light on the plausibility of this model, and thus indirectly on the plausibility of rank preservation, by examining test score transitions from baseline (start of P1) to endline (end of P3). For example, under the single factor model without measurement error, students in a given treatment arm in the top quartile of baseline scores should also end up in the top quartile of endline scores. Again, adding some noise to the test makes this prediction an approximate one, but we would want to see a relatively high transition probability, say 0.8 or 0.9, to support an "approximate rank preservation" interpretation.

Figure 2 plots the test score transitions within treatment arm by quartile, though the high fraction of students with zero baseline scores (because they have no background in reading when they arrive at school) forces us to combine the bottom two quartiles. More precisely, the figures show the probability of ending up in the upper quartile of the endline score distribution conditional on the student's quartile of the baseline score distribution.[47] The figure reveals that, while students who start out in the top quartile do have a higher probability of ending up in the top quartile within their treatment arm in all three arms, their advantage is quite modest. The same finding holds for the third quartile. Overall, the evidence in Figure 2 indicates either the failure of the one-factor model, a very noisy test, or both. We have a high degree of faith in the EGRA as a measure of basic reading ability, and so the latter theory seems unlikely.

The covariate balance at quantiles tests provide statistical evidence against rank preservation, though the relatively modest magnitudes of the estimated imbalances would support a view that rank preservation represents a rough approximation. The test score transition graphs in Figure 2, though, dissuade us from adopting that view. Instead, we interpret the QTEs solely as informing us about the effects of the NULP program variants on the distribution of outcomes, not as indicative of effects on students at a specific quantile of the

[46] Tests that omit the stratification cell fixed effects appear in Online Appendix Table 6. The qualitative results do not change.

[47] Note that Figure 2 uses only those students with non-missing values of baseline test scores.

status quo test score distribution.

[Figure 2 about here.]

[Table 4 about here.]

7 Systematic treatment effect variation

Having provided strong evidence of meaningful essential heterogeneity in our context, and having examined whether and what we can learn about that heterogeneity by considering additional substantive assumptions in the form of MSI and rank preservation, we now investigate the extent to which the treatment effect heterogeneity we observe correlates with observed covariates. We follow (some of) the literature in calling all of these variables "moderators".[48] Moreover, we follow Djebbari and Smith (2008) in dividing the extant treatment effect heterogeneity into a "systematic" component—the part that the moderators capture—and an "idiosyncratic" component—the part that the moderators do not capture, while keeping in mind that this division depends on the set of available candidate moderators.

Measuring and identifying systematic treatment effect heterogeneity adds value on several dimensions.[49] In the broader literature in education and labor economics, interest centers on using knowledge about "what works for whom" to target interventions on those most likely to benefit from them.[50] In the context of the NULP intervention, such targeting could only occur at the school or teacher level. Given the strong average effects of both versions of the NULP program, learning about predictors of relatively low (or even negative) treatment effects could also allow compensatory action within classrooms and could motivate further study of particular aspects of program implementation, both with an eye towards improving the treatment effects of those who presently benefit the least. Similarly, teachers whose characteristics predict lower average treatment effects could receive further training in program execution. Systematic variation may also shed light on how programs work, to the extent that theory and/or existing evidence associate specific causal mechanisms with specific moderators. Finally, in many contexts, policymakers care about particular groups

[48] We do not examine mediators, which the literature defines as intermediate outcomes that reflect particular causal pathways. In a sense, though, our test score outcome itself represents a mediator on the path to the adult outcomes that we really care about. Interpreted that way, we investigate moderators for our mediator.

[49] In addition to works cited elsewhere, on systematic heterogeneity see, e.g. Bitler, Gelbach, and Hoynes (2017), Lee and Shaikh (2014), and Weiss, Bloom, and Brock (2014).

[50] See, e.g. Berger, Black, and Smith (2001) in the context of active labor market programs.

for broader reasons, as with girls or ethnic minorities in primary and secondary school in much of the developing world.

7.1 Candidate Moderators

The design of the NULP evaluation did not have effect moderation as a primary goal. As a result, we lack data on many plausible moderators—see Section 7.4 for our wish list—and we lack the statistical power to detect modest but substantively meaningful moderators.[51] We group the candidate moderators into three sets: student characteristics, teacher characteristics, and school characteristics.

Theory and/or existing empirical evidence make the case for several of our candidate moderators. For example, models of education production like those in Hanushek (1992) and Todd and Wolpin (2003) often imply that the productivity of additional inputs depends on previous investments (so-called "dynamic complementarity"), and extensive research has shown that students' initial levels of preparation matter a great deal for their learning gains (Banerjee et al. 2016). This motivates our inclusion of baseline test scores. Claims that starting school later improves school performance, as in Gladwell (2008), suggest including student age among our candidate moderators. A large literature (mostly in the developed world) surveyed in Hanushek and Rivkin (2010) finds that teacher experience predicts teacher quality as measured by value-added. Buhl-Wiggers et al. (2018b) show that it does so for the teachers in our study too. As demonstrated by Angrist and Lavy (1999), and many others, at the school level, student-to-teacher ratios predict student learning. They also presumably affect how well teachers can implement interventions.[52] School size may capture economies of scale.

Policy interest drives the inclusion of some other candidate moderators. There is considerable demand for evidence on interventions that work well for girls—see, e.g. Evans and Yuan (2019)—as well as evidence in Lim and Meer (2017) that assigning girls to female teachers improves their test scores. These factors help motivate the inclusion of student sex and teacher sex. Easily-used measures to guide the assignment of NULP to particular schools (when funds do not allow universal implementation) would aid policymakers and educational administrators. For this reason, we include the PLE pass rate as a candidate moderator. Both administrators and parents commonly use it as a proxy for the quality of Ugandan primary schools and it is routinely collected and thus readily available.

[51] On this point see, e.g. Gelman (2018).

[52] Indeed, the initial phase of the NULP experiment, conducted on a separate cohort of students in 2013, imposed class-size restrictions (by requiring at least two teachers per grade) for exactly this reason (Kerwin and Thornton 2020).

Finally, practical considerations also affect out choices regarding candidate moderators. As noted above, we include an indicator for missing baseline test scores (and set missing scores to zero) because a large fraction of our student sample has no baseline data. We include teacher age and education levels because they strongly correlate with experience and might otherwise act as omitted confounders. More broadly, we do not include every potential moderator in the data in our set of candidate moderators. Instead, we omit many potential variables (ranging from the composition of students' households to teacher income) on a priori grounds in order to avoid over-fitting and conserve degrees of freedom in the conventional approach, and to avoid computational burden in the machine learning analysis.[53] One important criterion for these a priori omissions concerns item non-response; with the exception of baseline test scores, we only included variables with valid values for a large fraction of students in the study so that we could keep the sample size up without adding additional indicators for missing values.

7.2 Conventional estimates of systematic variation

7.2.1 Introduction

What we call the conventional approach simply takes some available moderators and includes them in the experimental impact linear regression model both as main effects and interacted with the treatment indicators. Depending on the available sample size and the size and nature of the set of candidate moderators, the set of included moderators may include all available candidate moderators, or some subset chosen in an ad hoc manner to avoid multicollinearity and/or over-fitting and/or concerns about multiple hypothesis testing. We can write the resulting linear model as

$$Y_{isc} = \beta_{FC}FC_s + \beta_{RC}RC_s + \sum_{j=1}^{J}[\beta_{FC}^{j}FC_{is}X_i^j + \beta_{RC}^{j}RC_{is}X_i^j] + \gamma^j X_i^j] + \beta_Y Y_{i,P1} + \beta_D D_i + \alpha_c + \epsilon_{isc} \tag{7}$$

As above, Y_{isc} denotes the outcome variable for student i in school s and in stratification cell c. FC_s and RC_s indicate assignment to the full-cost or reduced-cost treatment arm, respectively, with associated coefficients β_{FC} and β_{RC}. We let X_i^j denote the value of moderator $j \in 1, \ldots, J$ for student i, and we de-mean all the moderators prior to inclusion so

[53] We assess the underlying dimensionality of our set of candidate moderators via a principal components analysis. The results, shown in Online Appendix Table 7, reveal that the moderators do not, for the most part, measure overlapping constructs. Indeed, the most important component explains just 17% of the overall variance.

that β_{FC} and β_{RC} retain their interpretation as estimates of the average treatment effect. We include baseline scores $Y_{(i,P1)}$ (and an indicator for missing baseline scores) among the moderators in (7). The coefficients β^j_{FC} and β^j_{RC} indicate the conditional expected change in the relevant treatment effect for a one-unit change in the moderator, while γ^j indicates the conditional expected change in the untreated outcome for a one-unit change in moderator j.

Of course, while we randomly assigned the NULP treatments, we did not randomly assign the moderators. This immediately implies no causal interpretation of the γ^j without some explicit argument for an alternative source of identification—as in any non-experimental analysis. Though you would not know it from reading most moderation analyses using experimental data, the same point applies to the β^j_{FC} and β^j_{RC}. For example, if X^1_i indicates female rather than male students, a substantively large, positive, and statistically significant coefficient could imply that the treatment effect of NULP increases with some student characteristic that female students have more of than male students (and that does not appear among the remaining moderators) rather than that being female causes a higher treatment effect. We interpret our estimates accordingly, both here and in Section 7.3; see e.g. Hotz, Imbens, and Mortimer (2005) for further discussion. Finally, α_c is a treatment stratification cell fixed effect and, as always, ϵ_{isc} is a mean-zero term that captures the effects of all omitted determinants of test scores. We cluster the standard errors at the school level given the school-level treatment assignment.

7.2.2 Findings

In Table 5 we present the results of the conventional analysis of systematic treatment effect heterogeneity. Column (1) presents the base model without moderators—i.e. the same model as in column (4) of Table 1. We then present, in turn, specifications that interact the treatment indicator with student characteristics in column (2), with teacher characteristics in column (3), and with school characteristics in column (4). Column (5), our preferred specification, includes all three sets of candidate moderators.

We find only limited evidence of systematic variation in treatment effects. We do see that students with missing baseline scores tend to have smaller treatment effects, as do students in schools with more teachers. Presumably neither represents a causal moderation effect but instead both represent proxies for other aspects of the student, in the case of the missing baseline scores, and of the school, in the case of more teachers. Consistent with the limited predictive power of these interaction terms, the adjusted R-squared barely budges when we add them all to the model in column (5), rising from 0.170 to 0.188, or by about 10 percent.

As an alternative metric for the success of our candidate moderators at capturing systematic treatment effect variation within the context of the linear model in (7), we examine

the extent to which removing the variation they capture reduces the FH lower bound on the impact variance. We do this by generating adjusted versions of the outcome variable that subtract off the estimated interaction terms in (7), so that:

$$\tilde{Y}_{isc} = Y_{isc} - \sum_{j=1}^{J} [\hat{\beta}^{j}_{FC} FC_{is} X^{j}_{i} + \hat{\beta}^{j}_{RC} RC_{is} X^{j}_{i}] \tag{8}$$

We then reconstruct the FH bounds as above but using $\tilde{Y}_{isc}$ as the outcome variable in place of Y_{isc} Online Appendix Table 8 show the results. This metric confirms the message from Table 5: the new FH lower bounds on the impact standard deviation equal 1.08 SDs for the full-cost program and 0.66 from the reduced-cost program. In both cases these represent slight increases from the original values in Table 2, presumably due to noise introduced by the imprecision of the coefficient estimates on the interaction terms in (7).

[Table 5 about here.]

7.3 Machine-learning estimates of systematic variation

7.3.1 Introduction

Given the limited success of the conventional approach to capturing systematic treatment effect heterogeneity, we turn to an alternative approach based on algorithmic model selection or, as the young people say, Machine Learning (ML).[54] Algorithmic model selection has a long history in statistics; for example, Linhart and Zucchini (1986) review the large literature already in place over three decades ago.[55] Economists of that era tended to mock early ML methods like stepwise regression as delegating the thinking to the computer; a (largely) generational shift in attitudes away from that view has coincided with the rising prominence of ML in economics as documented in, e.g. Athey (2019).

ML has several advantages for the examination of systematic treatment effect heterogeneity relative to the conventional approach we applied in Section 7.2. First, it allows for an exhaustive model search across a space defined by the researcher. Second, it reduces the number of what Simmons, Nelson, and Simonsohn (2011) call "researcher degrees of freedom"; by automating the model-selection process, ML methods tie researchers' hands

[54] Old people think "ML" denotes "Maximum Likelihood".

[55] Even some "modern" machine learning methods go back farther than one might think from reading the current literature. To pick two examples familiar to us: Heckman et al. (1998) use Classification and Regression Tree (CART) methods and Black and Smith (2004) apply cross-validation in model selection.

and prevent them from "cherry picking" results they like. Third, newer ML methods address problems related to over-fitting and post-model-selection inference.[56,57] Despite these advantages, ML methods cannot improve on the set of available candidate moderators. In the context of systematic treatment effect variation, this means that ML methods can help locate moderators out of an existing list of variables and can (depending on the method and on the researcher's inputs) find important non-linearities and interactions among the candidate moderators.[58]

The literature offers two broad categories of ML techniques for systematizing treatment effect heterogeneity.[59] The first builds on the Least Absolute Selection and Shrinkage Operator (LASSO) estimator, which adds a penalty function in the sum of the absolute values of the coefficient estimates to the standard Ordinary Least Squares (OLS) objective function. Relative to OLS, the regularization implicit in the LASSO pushes coefficients toward zero, which avoids over-fitting in contexts with many candidate moderators.[60] See, e.g., Chen et al. (2017), Imai and Ratkovic (2013), Knaus, Lechner, and Strittmatter (2020), and Tian et al. (2014), for more detail on the LASSO and empirical applications in different substantive domains.

The second ("arboreal") category comprises variants of moderator selection algorithms based on regression trees. In this context, regression trees build on the intuitive idea of splitting the sample based on the values of particular moderators according to some criterion related to the amount of treatment effect heterogeneity obtained. For example, the algorithm might split our sample based on treatment-control difference in endline score by whether the baseline test score was above and below the median. A sequence of repeated splits forms a regression tree, wherein each leaf contains observations with a unique set of choices at the splits that define that tree. A set of such trees, with the order of the candidate moderators used to perform the splits randomized among the trees, constitutes a random forest. See, e.g., Wager and Athey (2018), Davis and Heller (2017), Foster, Taylor, and Ruberg (2011), Green and Kern (2012), and Hill (2011), and Hill and Su (2013).

[56] Guggenberger (2010) describes a similar post-model-selection inference problem in using a Durbin-Wu-Hausman test to choose whether to report OLS or IV estimates.

[57] Though not relevant in our setting, modern machine learning techniques also make easy work of situations with more candidate moderators (or, more generally, predictors) than observations. The conventional approach has no way to deal with such situations other than ruling out many candidate moderators on *a priori* grounds.

[58] Indeed, many researchers seem to have an astounding degree of optimism regarding the existence of heretofore undiscovered and substantively important third- and fourth-order interactions among moderators.

[59] James et al. (2017) provide an excellent textbook treatment of ML methods.

[60] Philosophically, one can either think of a world with many true zero coefficients (a world of "sparsity" in the jargon of ML), which the LASSO aims to find, or a world with many small but non-zero coefficients, which the LASSO approximates with zeros in finite samples. Though we have no real way to tell in which world we reside, it turns out to matter for the asymptotic theory.

We focus on LASSO-type methods in this paper for comparability to the conventional approach, which relies on linear regressions. In particular, we implement the method in Knaus, Lechner, and Strittmatter [hereinafter KLS] (2020), which struck us as particularly thoughtful.

7.3.2 Details of the machine-learning algorithm

The KLS (2020) approach combines the Modified Covariate Method (MCM) with the LASSO. MCM removes the main effects so that the LASSO implicitly considers only the interactions between the treatments and the moderators. We hand the LASSO the treatment indicators, plus the moderators described in Section 7.1. For comparison, we compare two covariate sets based on our candidate moderators. One, the "full" covariate set, includes the natural log and a fourth order polynomial of each (continuous) candidate moderator as well as all possible first-order interactions, and another (the "restricted" covariate set) omits the logs as well as the higher order terms and the interactions.[61] We choose the penalty parameter for the LASSO by 10-fold cross-validation to minimize post-LASSO mean-squared error.[62] Following KLS (2020), we conduct "honest" inference by splitting the sample in half at random, selecting the model on one half of the sample, and using the remaining half to estimate the treatment effects. Following the literature, we do this for 30 random splits of the sample and then, rather than averaging, we describe the overall patterns of estimates that emerge from the exercise.

7.3.3 Findings

Table 6 presents estimates from the specifications chosen by the LASSO for two of the 30 random sample splits for the full covariate set, in columns (1) and (2), and for the restricted set, in columns (3) and (4). For each covariate set, we present the two specifications with the median adjusted R-squared values. The restricted covariate set specifications reveal several patterns of interest: First, the LASSO drops many variables, but not that many. Of the 24 interactions between the full-cost and reduced-cost indicators and the candidate moderators included in the conventional model in Table 5, the specification in column (3) and (4) keep 15—more than we expected given the mushy findings in Table 5. Second, for both specifications, dropping nine variables reduces the adjusted R-squared only marginally, from 0.188 to 0.186, as the LASSO focuses attention on stronger predictors of treatment effects. Third, the coefficients on many of the variables kept by the LASSO do not obtain

[61] For the full covariate set model we discretize the baseline test score in order to avoid extreme outliers associated with higher order polynomial terms.

[62] Online Appendix Table 9 shows the values chosen for each sample split.

traditional levels of statistical significance; the LASSO objective function differs from that of other algorithms such as stepwise regression. Fourth, of the variables attaining statistical significance in Table 5, both the column (3) and column (4) specifications retain the missing baseline test score moderator for the full-cost program while neither retains the number of teachers as a moderator for the reduced-cost program. Finally, as one would expect from our moderate sample sizes and the non-trivial correlations among the candidate moderators, the LASSO retains somewhat different moderators in the two sample splits. The right-hand column of Table 7 presents a "league table" with the top 20 candidate moderators in order by the number of sample splits in which they find favor with the LASSO.

Comparing the specifications for the full covariate set to those for the restricted covariate set in Table 6 yields some additional findings. First, the additional covariates matter in the sense that they noticeably increase the median adjusted R-squared, from 0.186 to 0.213. Second, many of the simple treatment-moderator interactions kept by the LASSO in the case of the reduced covariate set get dropped by the LASSO in the full covariate set in favor of more complicated interactions involving more than one moderator. Third, the LASSO does not retain a single cubic or quartic term in any of the continuous moderators in these specifications (though a couple of cubic terms do get retained in most other sample splits). Fourth, despite the support for it in the literature, none of the 30 sample splits provoke the LASSO to retain the interaction between the sex of the student and the sex of the teacher. Finally, it bears repeating that which variables get retained varies among the sample splits; the reader should take most seriously those interactions that end up near the top of the rankings in Table 7.

Given the complexities of the patterns in the selected predictors across models, we follow KLS (2020) and focus on the patterns of Conditional Average Treatment Effects (CATEs) across all 30 sample splits. The estimated CATE for each unit from a given sample split equals the predicted treatment effect based on the parameter estimates from the specification selected by the LASSO for that split—essentially a predicted value from the analogue of equation (7) for a given specification. We then average the CATEs for each student over the 30 sample splits to produce what we call ACATEs (i.e. "aggregated" CATEs, following KLS (2020)), and examine how they vary across students and correlate with student characteristics.

Figure 3 plots kernel densities of the estimated ACATEs for the full covariate set for both the full- and reduced-cost programs; Online Appendix Figure 1 does the same for the restricted covariate set.[63] Online Appendix Table 10 provides the corresponding descriptive

[63] We do not "shrink" the estimates to account for sampling variation; at the same time, averaging over 30 CATEs limits the sampling variation they contain.

statistics. In Figure 3, we find a much higher variance for the full-cost program than for the reduced-cost program; both distributions center near the corresponding ATE. Not surprisingly, we also obtain substantially more variable ACATEs from the full covariate set than from the restricted covariate set.[64] Despite all this variation in the ACATEs, subtracting them from the outcomes as in equation (8) and repeating the FH bounding exercise again barely moves the estimated lower bound on the impact variation, which declines from 1.066 to 1.021 (about four percent) for the full-cost program and rises from 0.642 to 0.648 for the reduced-cost program (Online Appendix Table 11).[65]

Figure 4 illustrates how the ACATEs vary with a subset of our moderators. We split each variable into "high" and "low" ranges where "high" denotes above the median for continuous variables and a value of one for indicators; "low" is the complement of "high". Some interesting differences emerge—for example, teachers with more years of education produce lower CATEs for the full-cost treatment—but few attain statistical significance. The noteworthy exceptions arise for the full-cost program: baseline test scores correlate positively with ACATEs while the indicator for a missing baseline scores correlates negatively with them.

Table 8 does the reverse of Figure 4 and examines the mean value of particular moderators conditional on the sign of the ACATE. We find many statistically meaningful differences: students with positive ACATEs have higher test scores, are more likely to have a male teacher, and are more likely to have an inexperienced teacher. For the full-cost program, students with positive ACATEs are more likely to be male; for the reduced-cost version they are in schools with more pupils per teacher.[66]

[Figure 3 about here.]

[Figure 4 about here.]

[Table 6 about here.]

[64] Online Appendix Figure 2 shows the estimated density of the ACATEs if we do not discretize the baseline test scores. Individual students with very high and low CATEs due to the inclusion of quadratic and higher-order terms in the baseline test score as a moderator in particular specifications drive the long upper and lower tails.

[65] This small increase again reflects a combination of the limited explanatory power of the moderators and sampling variation.

[66] The apparent inconsistency between Figure 4 and Table 8 results from differences in the nature and extent of the underlying conditional variation. Underlying Figure 4, we have considerable variation in the ACATEs conditional on the value of specific moderators. Conversely, holding the ACATEs fixed leaves relatively little variation in specific moderators. This pattern is partly mechanical in the sense that the ACATEs can and do vary widely while many of the specific moderators have a low, bounded variance.

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

7.4 Could "better data help a lot"?

The overarching conclusion from the analyses in the two preceding sections is surely that we do a very poor job indeed of converting the treatment effect heterogeneity we know exists from the FH bounds into systematic heterogeneity. Instead, but for a tiny fraction, it remains stubbornly idiosyncratic. We see two broad potential conclusions from this finding, one pessimistic and one optimistic. The pessimistic one sees the heterogeneity as practically irreducible, i.e. that the important moderators lie outside the bounds of what social scientists can effectively measure at scale. The (relatively) optimistic one sees it as a pointed reminder that we (and by "we" we mean the literature in general, not just our study) have not really pushed that hard on either the theory or the measurement of effect moderation in the context of educational interventions, especially in a developing country context. At the margin, more effort on theory and data collection related to moderators might have a substantially higher knowledge payoff than the same amount of time and effort devoted to developing the 12,343rd tweak of the LASSO.

To distinguish between these two views, future research on related interventions should collect data on new and different moderators. One way to come up with new moderators builds on what little we already do know. We find, and others find, that baseline test scores predict treatment effects. Collecting additional baseline exam scores, or more measures of baseline cognitive skills in general, could reduce the error and/or increase the dimensionality with which we measure the underlying construct of student ability. The fact that the LASSO quite often retains moderators that include school-wide student counts and multi-grade-level pupil-teacher numbers implies that actual class sizes—which are not available on a consistent basis in our sample—merit examination. Along similar lines, the presence of third-grade teacher characteristics among the commonly retained candidate moderators signals the value of looking at the characteristics of teachers from earlier grades and of collecting additional characteristics (e.g. test scores, soft skills) for all teachers.

Another way of thinking about useful moderators imagines Holmesian "dogs that didn't bark": sets of variables completely absent from our current data. A leading candidate is students' non-cognitive skills, e.g. the sorts of "soft skills" considered by Heckman and

Kautz (2012). We also lack data on family characteristics such as parental education or books in the house, on parental investments prior to the initiation of schooling, and on pre-natal (or even post-natal) environmental exposures. Along similar lines, we have no direct measures of pre-intervention teaching quality, such as a value-added score or head teacher evaluation. Buhl-Wiggers et al. (2018b) show that the NULP program shifts the distribution of teacher value-added, suggesting potentially important interactions between teacher quality and the program's effects.

While all the moderators on our list lie well within the capacities of social science measurement, they do not, unlike nearly all of our current candidates, commonly appear in administrative data systems, particularly in a developing country context. We do not expect such systems to capture non-cognitive skills or teacher value-added anytime soon, let alone parental characteristics or early childhood investments and shocks. Thus, the knowledge gained from studying these additional moderators will likely not have much immediate value for the day-to-day implementation of interventions like NULP. More broadly, these and other potential moderators might well perform no better than our actual ones, in which case uncertainty about the student-specific effects of education interventions may simply be a fact of life.

8 Conclusion

Using data from a randomized evaluation of a highly effective literacy program in 128 primary schools in northern Uganda, we show that the program's effects vary widely across students. We resoundingly reject the null hypothesis of equal student-level effects. For the full-cost version of the program, the FH lower bound on the impact standard deviation exceeds 1.0 SD of our endline Leblango reading test score index. This implies that the variation in gains within this program is larger than the difference in the mean effects across the two versions implemented in our study. Indeed, there is more variation in student-level gains within this one program than in the mean treatment effects of all developing-country primary education programs ever studied in randomized trials. At the same time, the full-cost program's average gain of 1.4 SDs masks the fact that, assuming normally distributed treatment effects, at least 29% of students experience a gain of more than 2 SDs, while the program makes more than 10% worse off.

Who exactly benefits from the intervention, and who gets left behind? We use various techniques to try to answer this question, with remarkably little success. Imposing a stochastic increasingness assumption as in FL (2020) concentrates any possible negative mean effects at the upper end of the outcome distribution but otherwise delivers disappoint-

ingly wide bounds on the expected effects for students at particular quantiles. Traditional quantile treatment effect estimates imply much bigger increases at the top of the distribution than at the bottom—but we can easily reject the assumption of rank preservation, and thus our QTEs do not tell us about the gains for students at a given quantile of the *status quo* distribution. Finally, both our conventional linear moderation analysis and our application of modern ML methods fail to induce our set of available moderators to explain much of the underlying variation in impacts.

Our results leave unanswered the question of exactly why this intervention leaves some children behind. Following Pritchett and Beatty (2015), one candidate explanation argues that instructional methods should better reflect student ability levels.[67] Even though the NULP model begins with the basics of reading and intentionally goes slower than the status quo literacy lessons, it may still move too quickly for some students. In particular, given that this program begins upon students' entry into the school system, certain students could lack foundational skills needed for literacy acquisition. This line of reasoning suggests that tracking students by ability might add value even in the context of a program whose untracked version has large average effects.[68]

We draw three major conclusions from our findings. First, identifying interventions that work on average will not fully address the "learning crisis" in developing countries. Discussions of the learning crisis emphasize that even though school enrollments have risen in the developing world, many students end up learning nothing (World Bank 2018). We find exactly this pattern in the highly effective intervention we study in this paper. Whether such interventions represent good public policy, however, rests on the shape of the returns to education and also on normative judgments. If education exhibits convex returns, then the best investments may boost the upper end of the performance distribution—as our quantile treatment effects analyses show happens with the NULP. Even in that case, ethical or political conditions may push against running education systems in ways that help some students while leaving others behind.

Second, we find clear evidence of statistically and substantively meaningful variation in the treatment effects for yet another program category in yet another context. Nonetheless, despite several decades of evidence, reporting basic non-parametric estimates of the lower bound on the variation in treatment effects remains rare in program evaluations. In our view, reporting these bounds should become standard practice for future randomized trials

[67] This idea sits at the core of the "Teaching at the Right Level" program in Banerjee et al. (2016); the US-based Response to Intervention method also targets interventions by student performance levels; see, e.g. Mesmer and Mesmer (2008).

[68] Duflo, Dupas, and Kremer (2011) provide an example of the effectiveness of tracking in a developing country.

in education as well as other domains. Furthermore, studies that examine systematic treatment effect heterogeneity should report how the lower bound changes when the estimated systematic heterogeneity is removed.

Third, our set of "usual suspects" moderators capture very little in the way of systematic treatment effect heterogeneity, even when exploited by a state-of-the art ML algorithm. While we attribute some part of this failure to our modest sample size, we assign the bulk of it to a general failure in the literature to push forward with the applied theory of effect moderation in education interventions and with the measurement of existing but heretofore unexamined potential moderators. We think that better data would yield higher returns at the margin than further refinements to existing ML methods.

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002). "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings". *Econometrica* 70.1, pp. 91–117.

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. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2009.03.004.

Angrist, Joshua D. and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". *The Quarterly Journal of Economics* 114.2, pp. 533–575. ISSN: 0033-5533. DOI: 10.1162/003355399556061.

Athey, Susan (2019). "The Impact of Machine Learning on Economics". *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, p. 31.

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

Berger, Mark C., Dan Black, and Jeffrey A. Smith (2001). "Evaluating Profiling as a Means of Allocating Government Services". *Econometric Evaluation of Labour Market Policies*. Springer, pp. 59–84.

Bhattacharya, Jay, Azeem M. Shaikh, and Edward Vytlacil (2008). "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization". *The American Economic Review* 98.2, pp. 351–356. ISSN: 0002-8282.

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

— (2006). "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments". *The American Economic Review* 96.4, pp. 988–1012.

— (2017). "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment". *The Review of Economics and Statistics*. ISSN: 0034-6535.

Bitler, Marianne P., Hilary W. Hoynes, and Thurston Domina (2016). "Experimental Evidence on Distributional Effects of Head Start". *Working Paper*.

Black, Dan A. and Jeffrey A. Smith (2004). "How Robust Is the Evidence on the Effects of College Quality? Evidence from Matching". *Journal of Econometrics*. Higher education (Annals issue) 121.1, pp. 99–124. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2003.10.006.

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.

Boone, Peter, Ila Fazzio, Kameshwari Jandhyala, Chitra Jayanty, Gangadhar Jayanty, Simon Johnson, Vimala Ramachandrin, Filipa Silva, and Zhaoguo Zhan (2013). *The Surprisingly Dire Situation of Children's Education in Rural West Africa: Results from the CREO Study in Guinea-Bissau (Comprehensive Review of Education Outcomes)*. w18971. National Bureau of Economic Research.

Bruhn, Miriam and David McKenzie (2009). "In Pursuit of Balance: Randomization in Practice in Development Field Experiments". *American Economic Journal: Applied Economics* 1.4, pp. 200–232. ISSN: 1945-7782.

Brunette, Tracy, Benjamin Piper, Rachel Jordan, Simon King, and Rehemah Nabacwa (2019). "The Impact of Mother Tongue Reading Instruction in Twelve Ugandan Languages and the Role of Language Complexity, Socioeconomic Factors, and Program Implementation". *Comparative Education Review* 63.4, pp. 591–612. ISSN: 0010-4086. DOI: 10.1086/705426.

Buhl-Wiggers, Julie, Jason Kerwin, Jeffrey Smith, and Rebecca Thornton (2018a). *Program Scale-up and Sustainability*. Working Paper.

— (2018b). "Teacher Effectiveness in Africa: Longitudinal and Causal Estimates". *Working Paper* S-89238-UGA-1.

Cambanis, Stamatis, Gordon Simons, and William Stout (1976). "Inequalities for E(k(X,Y)) when the marginals are fixed". *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 36, pp. 285–294.

Chen, Shuai, Lu Tian, Tianxi Cai, and Menggang Yu (2017). "A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring". *Biometrics* 73.4, pp. 1199–1209. ISSN: 1541-0420. DOI: 10.1111/biom.12676.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor (2019). "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and in-Classroom Coaching". *Journal of Human Resources* in press. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.55.3.0618-9538R1.

Conn, Katharine M. (2017). "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations". *Review of Educational Research* 87.5, pp. 863–898. DOI: 10.3102/0034654317712025.

Csörgo, Miklos (1983). *Quantile Processes with Statistical Applications*. Vol. 42. Siam.

Davis, Jonathan M. V. and Sara B. Heller (2017). "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs". *American Economic Review* 107.5, pp. 546–550. ISSN: 0002-8282. DOI: 10.1257/aer.p20171000.

Delavallade, Clara, Alan Griffith, and Rebecca Thornton (2019). *Effects of a Multi-Faceted Education Program on Enrollment, Equity, Learning, and School Management: Evidence from India*. Policy Research Working Paper 9081. Washington, DC: The World Bank.

Djebbari, Habiba and Jeffrey Smith (2008). "Heterogeneous Impacts in Progresa". *Journal of Econometrics* 145.1, pp. 64–80.

Doksum, Kjell (1974). "Empirical probability plots and statistical inference for nonlinear models in the two-sample case". *The Annals of Statistics*, pp. 267–277.

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

Evans, David K. and Anna Popova (2016). "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews". *The World Bank Research Observer* 31.2, pp. 242–270. ISSN: 0257-3032. DOI: 10.1093/wbro/lkw004.

Evans, David K. and Fei Yuan (2018). "Commentary on Chapter 11, 'Learning at the bottom of the pyramid' and the global targets in education". *Learning at the Bottom of the Pyramid: Science, Measurement, and Policy in Low-Income Countries*. Ed. by Daniel A. Wagner, Sharon Wolf, and Robert F. Boruch. Paris, France: International Institute for Educational Planning, pp. 232–233. ISBN: 978-92-803-1420-5.

— (2019). *What We Learn about Girls' Education from Interventions that Do Not Focus on Girls*. Policy Research Working Papers. The World Bank. DOI: 10.1596/1813-9450-8944.

Fan, Yanqin and Sang Soo Park (2010). "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference". *Econometric Theory* 26.3, pp. 931–951. ISSN: 0266-4666, 1469-4360. DOI: 10.1017/S0266466609990168.

Foster, Jared C., Jeremy MG Taylor, and Stephen J. Ruberg (2011). "Subgroup Identification from Randomized Clinical Trial Data". *Statistics in Medicine* 30.24, pp. 2867–2880.

Frandsen, Brigham and Lars Lefgren (2020). "Partial Identification of the Distribution of Treatment Effects with an Application to the Knowledge Is Power Program (KIPP)". *Working Paper*.

Fréchet, M. (1951). "Les Tableaux de Corrélation Dont les Marges Sont Données". *Annales de l'Université de Lyon. Section A: Sciences, Mathématiques et Astronomie* 14, pp. 53–77.

Freedman, David A. (2008). "On Regression Adjustments to Experimental Data". *Advances in Applied Mathematics* 40.2, pp. 180–193.

Gelman, Andrew (2018). *You need 16 times the sample size to estimate an interaction than to estimate a main effect*. Statistical Modeling, Causal Inference, and Social Science. URL: https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/ (visited on 08/05/2020).

Gladwell, Malcolm (2008). *Outliers: The story of success*. Little, Brown.

Glewwe, Paul W., Eric A. Hanushek, Sarah D. Humpage, and Renato Ravina (2013). "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010". *Education Policy in Developing Countries*. Ed. by Paul W. Glewwe. doi:10.7208/chicago/9780226078854.003.0002. Chicago and London: University of Chicago Press.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin (2009). "Many Children Left Behind? Textbooks and Test Scores in Kenya". *American Economic Journal: Applied Economics* 1.1, pp. 112–135. ISSN: 1945-7782. DOI: 10.1257/app.1.1.112.

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.

Gove, Amber, Tracy Brunette, Jennae Bulat, Bidemi Carrol, Catherine Henny, Wykia Macon, Evangeline Nderu, and Yasmin Sitabkhan (2017). "Assessing the Impact of Early Learning Programs in Africa: Assessing the Impact of Early Learning Programs in Africa". *New Directions for Child and Adolescent Development* 2017.158, pp. 25–41. ISSN: 15203247. DOI: 10.1002/cad.20224.

Green, Donald P. and Holger L. Kern (2012). "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees". *Public Opinion Quarterly* 76.3, pp. 491–511.

Guggenberger, Patrik (2010). "The Impact of a Hausman Pretest on the Asymptotic Size of a Hypothesis Test". *Econometric Theory* 26.2, pp. 369–382. ISSN: 1469-4360, 0266-4666. DOI: 10.1017/S0266466609100026.

Hanushek, Eric A. (1992). "The Trade-off between Child Quantity and Quality". *Journal of Political Economy* 100.1, pp. 84–117. ISSN: 0022-3808.

Hanushek, Eric A. and Steven G. Rivkin (2010). "Generalizations about Using Value-added Measures of Teacher Quality". *American Economic Review* 100.2, pp. 267–71.

Heckman, J. J., J. Smith, and N. Clements (1997). "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts". *The Review of Economic Studies* 64.4, pp. 487–535. ISSN: 0034-6527, 1467-937X. DOI: 10.2307/2971729.

Heckman, James J. and Tim Kautz (2012). "Hard Evidence on Soft Skills". *Labour Economics*. European Association of Labour Economists 23rd annual conference, Paphos, Cyprus, 22-24th September 2011 19.4, pp. 451–464. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2012.05.014.

Heckman, James J, Sergio Urzua, and Edward Vytlacil (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity". *The Review of Economics and Statistics* 88.3, p. 58.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing Selection Bias Using Experimental Data". *Econometrica* 66.5, pp. 1017–1098. DOI: 10.2307/2999630.

Hill, Jennifer L. (2011). "Bayesian Nonparametric Modeling for Causal Inference". *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Hill, Jennifer and Yu-Sung Su (2013). "Assessing Lack of Common Support in Causal Inference using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Children's Cognitive Outcomes". *The Annals of Applied Statistics*, pp. 1386–1420.

Höffding, W. (1940). "Masstabinvariante Korrelationsmasse für Diskontinuierliche Verteilungen." *Arkiv für Matematischen Wirtschaften and Sozialforschung* 7, pp. 49–70.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer (2005). "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations". *Journal of Econometrics* 125.1, pp. 241–270.

Imai, Kosuke and Marc Ratkovic (2013). "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation". *The Annals of Applied Statistics* 7.1, pp. 443–470.

Jackson, Kirabo and Alexey Makarin (2018). "Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment". *American Economic Journal: Economic Policy* 10.3, pp. 226–254. ISSN: 1945-7731. DOI: 10.1257/pol.20170211.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2017). *An Introduction to Statistical Learning*. 8th. Springer. ISBN: 978-1-4614-7137-0.

Kayabwe, Samuel, Rehemah Nabacwa, Joseph Eilor, and Rosemary Waya Mugeni (2014). *The Use and Usefulness of School Grants: Lessons from Uganda*. IIEP Country Notes. Paris, France: International Institute for Educational Planning.

Kerwin, Jason T. and Rebecca L. Thornton (2020). "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures". *The Review of Economics and Statistics*, pp. 1–45. ISSN: 0034-6535. DOI: 10.1162/rest_a_00911.

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter (2020). "Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach". *Journal of Human Resources*, 0718–9615R1. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.57.2.0718-9615R1.

Koenker, Roger and Gilbert Bassett (1978). "Regression Quantiles". *Econometrica* 46.1, pp. 33–50.

Koenker, Roger and Yannis Bilias (2002). "Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments". *Economic Applications of Quantile Regression*. Springer, pp. 199–220.

Kremer, Michael, Conner Brannen, and Rachel Glennerster (2013). "The Challenge of Education and Learning in the Developing World". *Science* 340.6130, pp. 297–300.

Krishnaratne, Shari, Howard White, and Ella Carpenter (2013). *Quality Education for All Children? What Works in Education in Developing Countries*. Working Paper 20. New Delhi: International Initiative for Impact Evaluation (3ie).

Lee, Soohyung and Azeem M. Shaikh (2014). "Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of Progresa on School Enrollment: Heterogeneous Treatment Effects". *Journal of Applied Econometrics* 29.4, pp. 612–626. ISSN: 08837252. DOI: 10.1002/jae.2327.

Lehmann, Erich Leo and Howard J. D'Abrera (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day.

Lim, Jaegeum and Jonathan Meer (2017). "The Impact of Teacher–Student Gender Matches Random Assignment Evidence from South Korea". *Journal of Human Resources* 52.4, pp. 979–997. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.52.4.1215-7585R1.

Lin, Winston (2013). "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique". *Annals of Applied Statistics* 7.1, pp. 295–318. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/12-AOAS583.

Linhart, H. and W. Zucchini (1986). *Model Selection*. New York: Wiley. 301 pp. ISBN: 978-0-471-83722-0.

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.

Mesmer, Eric M. and Heidi Anne E. Mesmer (2008). "Response to Intervention (RTI): What Teachers of Reading Need to Know". *The Reading Teacher* 62.4, pp. 280–290. ISSN: 1936-2714. DOI: 10.1598/RT.62.4.1.

Ministry of Education and Sports (2016). *Uganda Educational Statistical Abstract*. Kampala, Uganda.

Murnane, Richard J. and Alejandro J. Ganimian (2014). *Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations*. Harvard University OpenScholar.

Piper, Benjamin (2010). *Uganda Early Grade Reading Assessment Findings Report: Literacy Acquisition and Mother Tongue*. Research Triangle Institute.

Piper, Benjamin, Stephanie S. Zuilkowski, and Salome Ong'ele (2016). "Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya". *Comparative Education Review* 60.4, pp. 776–807. ISSN: 0010-4086. DOI: 10.1086/688493.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Dunston Kwayumba, and Arbogast Oyanga (2018). "Examining the Secondary Effects of Mother-Tongue Literacy Instruction in Kenya: Impacts on Student Learning in English, Kiswahili, and Mathematics". *International Journal of Educational Development* 59, pp. 110–127. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2017.10.002.

Pritchett, Lant and Amanda Beatty (2015). "Slow Down, You're Going Too Fast: Matching Curricula to Student Skill Levels". *International Journal of Educational Development* 40, pp. 276–288. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2014.11.013.

Read, T. and S. Enyutu (2005). *Road Map for the Implementation of the Curriculum Reforms Recommended by the Primary Curriculum Review Report and Approved by the Ministry of Education and Sports*. Final revised version. Kampala.

Rossell, Christine H. and Keith Baker (1996). "The Educational Effectiveness of Bilingual Education". *Research in the Teaching of English* 30.1, pp. 7–74. ISSN: 0034-527X.

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

Rudalevige, Andrew (2003). "The Politics of No Child Left Behind". *Education Next* 3.4, pp. 63–69.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as

Significant". *Psychological Science* 22.11, pp. 1359–1366. ISSN: 0956-7976, 1467-9280. DOI: 10.1177/0956797611417632.

Smits, Jeron, Janine Huisman, and Karine Kruijff (2008). *Home Language and Education in the Developing World*. Background paper prepared for the Education for All Global Monitoring Report 2009 2009/ED/EFA/MRT/PI/21. UNESCO.

Ssentanda, Medadi (2013). "Thematic Curriculum and Mother Tongue Education in Uganda: Discrepancies Between De Jure and De Facto Language-in-Education Policy". Multidisciplinary Approaches in Language Policy and Planning Conference. University of Calgary.

Tchen, André H. (1980). "Inequalities for Distributions with Given Marginals". *The Annals of Probability* 8.4, pp. 814–827.

Tian, Lu, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani (2014). "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates". *Journal of the American Statistical Association* 109.508, pp. 1517–1532. ISSN: 0162-1459. DOI: 10.1080/01621459.2014.951443.

Todd, Petra E. and Kenneth I. Wolpin (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement". *The Economic Journal* 113.485, F3–F33. ISSN: 0013-0133.

Uganda Bureau of Statistics (2017). *The National Population and Housing Census 2014 – Education in the Thematic Report Series*. Kampala, Uganda.

UNESCO (2017). *A Guide for Ensuring Inclusion and Equity in Education*. ISBN 978-92-3-100222-9. Paris, France: United Nations Educational, Scientific and Cultural Organization.

Uwezo (2019). *Are Our Children Learning? Uwezo Uganda Eighth Learning Assessment Report*. Kampala: Twaweza East Africa.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113.523, pp. 1228–1242. ISSN: 0162-1459. DOI: 10.1080/01621459.2017.1319839.

Webley, K. (2006). *Mother Tongue First: Children's Right to Learn in their Own Languages*. id21. Development Research Reporting Service, UK.

Weiss, Michael J., Howard S. Bloom, and Thomas Brock (2014). "A Conceptual Framework for Studying the Sources of Variation in Program Effects". *Journal of Policy Analysis and Management* 33.3, pp. 778–808.

Williamson, Robert C. and Tom Downs (1990). "Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds". *International Journal of Approximate Reasoning* 4.2, pp. 89–158. ISSN: 0888-613X. DOI: 10.1016/0888-613X(90)90022-T.

World Bank (2018). *World Development Report 2018: Learning to Realize Education's Promise*. doi:10.1596/978-1-4648-1096-1. Washington, DC: World Bank.

Tables

Table 1
Average Treatment Effects

	(1)	(2)	(3)	(4)
Full-cost	1.444***	1.401***	1.526***	1.396***
	(0.136)	(0.116)	(0.125)	(0.116)
Reduced-cost	0.795***	0.738***	0.794***	0.738***
	(0.103)	(0.109)	(0.116)	(0.108)
Baseline Test Score				0.387***
				(0.060)
1(BL Missing)				-0.126**
				(0.051)
Raw Baseline Test Score			0.395***	
			(0.063)	
Observations	4,868	4,868	2,395	4,868
R-squared	0.125	0.166	0.219	0.179
Adj-R-Squared	0.124	0.158	0.203	0.170
Group*Year*Cohort FE		Yes	Yes	Yes

Notes: Columns (1), (2), and (4) use the main analysis sample. Column (3) uses the subset of the main analysis sample with non-missing baseline test score index. Outcome is the Leblango reading test score index, standardized with respect to the control group. Heteroskedasticity-robust standard errors, clustered by schools, in parentheses. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table 2
Fréchet-Höffding Bounds

	Full-cost Program		Reduced-cost Program	
	Rank Preservation (1)	Rank Inversion (2)	Rank Preservation (3)	Rank Inversion (4)
Percentiles under control status				
5^{th}	0.034	5.631	0.034	4.553
	(0.006)	(0.104)	(0.006)	(0.120)
25^{th}	0.386	3.318	0.205	2.199
	(0.056)	(0.167)	(0.025)	(0.137)
50^{th}	1.333	1.333	0.619	0.619
	(0.173)	(0.173)	(0.094)	(0.094)
75^{th}	2.577	-0.355	1.458	-0.536
	(0.188)	(0.104)	(0.166)	(0.092)
95^{th}	2.964	-2.633	1.886	-2.633
	(0.196)	(0.162)	(0.190)	(0.162)
Impact Standard Deviation	1.066	2.615	0.642	2.218
	(0.023)	(0.016)	(0.020)	(0.011)
Outcome Correlation	0.932	-0.655	0.975	-0.577
	(0.014)	(0.023)	(0.009)	(0.017)
Fraction Positive	0.980	0.697	0.980	0.646
	(0.005)	(0.015)	(0.006)	(0.017)

Notes: Estimates use the main analysis sample. Table presents Fréchet-Höffding bounds of the EGRA Leblango test score distribution, computed as described in Section 5.2. Standard errors computed using 1000 bootstrap replications.

Table 3
Frandsen and Lefgren Bounds on Treatment Effects by Percentile

	Full-cost Program		Reduced-cost Program	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)
Percentiles under control status				
5^{th}	0.195	2.652	0.163	1.919
	(0.057)	(0.057)	(0.065)	(0.047)
25^{th}	0.307	3.015	0.183	2.160
	(0.029)	(0.071)	(0.039)	(0.058)
50^{th}	0.565	3.774	0.263	2.732
	(0.049)	(0.092)	(0.030)	(0.083)
75^{th}	0.600	4.269	0.085	3.140
	(0.052)	(0.104)	(0.032)	(0.100)
95^{th}	-0.573	4.246	-1.193	3.233
	(0.050)	(0.164)	(0.039)	(0.185)

Notes: Estimates use the main analysis sample. School-clustered bootstrap standard errors computed using 100 replications in parentheses.

Table 4
Tests of Covariate Balance by Endline Test Score Quartile

	Full-Cost				Reduced-Cost			
	0-25th Perc.	25-50th Perc.	50-75th Perc.	75-100th Perc.	0-25th Perc.	25-50th Perc.	50-75th Perc.	75-100th Perc.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline Test Score	-0.008	-0.005	0.021	0.008	-0.015	0.019	-0.008	0.030
	[-0.028;0.030]	[-0.032;0.036]	[-0.041;0.040]	[-0.081;0.075]	[-0.029;0.031]	[-0.034;0.032]	[-0.039;0.042]	[-0.079;0.072]
1(BL Missing)	0.002	-0.065*	0.043	0.040	0.024	-0.079**	0.016	0.037
	[-0.056;0.057]	[-0.057;0.058]	[-0.060;0.059]	[-0.056;0.061]	[-0.055;0.051]	[-0.056;0.062]	[-0.056;0.060]	[-0.053;0.057]
1(Student Male)	0.028	-0.054	-0.027	0.043	0.077**	-0.018	-0.076**	0.053
	[-0.063;0.060]	[-0.063;0.058]	[-0.056;0.057]	[-0.057;0.059]	[-0.057;0.059]	[-0.063;0.058]	[-0.058;0.054]	[-0.054;0.055]
Student's Age	-0.025	-0.114	0.179**	-0.046	0.002	-0.131	0.107	-0.018
	[-0.141;0.140]	[-0.147;0.140]	[-0.144;0.121]	[-0.127;0.116]	[-0.141;0.142]	[-0.135;0.128]	[-0.130;0.123]	[-0.118;0.118]
1(Male Teacher)	-0.091***	-0.024	0.012	0.048**	-0.002	-0.004	-0.035	0.005
	[-0.039;0.038]	[-0.039;0.042]	[-0.044;0.040]	[-0.036;0.035]	[-0.036;0.040]	[-0.038;0.042]	[-0.042;0.042]	[-0.035;0.034]
Teacher's Age	-1.077***	0.216	-0.048	0.549*	-2.470***	0.024	0.729*	0.289
	[-0.611;0.635]	[-0.657;0.681]	[-0.656;0.662]	[-0.557;0.540]	[-0.605;0.577]	[-0.638;0.659]	[-0.637;0.637]	[-0.540;0.553]
Teacher's Experience	0.063	-0.090	-0.269	-0.310	-0.974***	-0.120	0.507	-0.334
	[-0.557;0.536]	[-0.542;0.541]	[-0.536;0.591]	[-0.503;0.486]	[-0.540;0.524]	[-0.524;0.532]	[-0.542;0.578]	[-0.508;0.472]
Years of Education	-0.145***	0.027	0.013	0.116**	-0.081	0.013	0.098	-0.031
	[-0.086;0.079]	[-0.101;0.098]	[-0.100;0.098]	[-0.096;0.097]	[-0.081;0.082]	[-0.103;0.104]	[-0.092;0.104]	[-0.098;0.095]
School's Enrollment	-5.341	-9.405	-4.079	-5.977	17.772**	4.160	-1.445	-16.225*
	[-15.418;14.343]	[-17.243;16.805]	[-17.296;17.480]	[-15.159;13.392]	[-15.175;14.185]	[-16.598;18.389]	[-19.085;16.757]	[-15.007;14.078]
Pupil-Teacher-Ratio	1.769**	0.767	-3.431***	-1.172*	-0.511	-0.178	-0.806	0.741
	[-1.182;1.002]	[-1.190;1.168]	[-1.316;1.330]	[-0.999;0.914]	[-1.106;1.051]	[-1.247;1.307]	[-1.374;1.237]	[-1.015;0.971]
PLE Pass Rate	0.003***	0.001	0.002	-0.003***	0.007***	0.001	-0.000	-0.003***
	[-0.002;0.001]	[-0.002;0.001]	[-0.002;0.002]	[-0.001;0.001]	[-0.002;0.002]	[-0.002;0.001]	[-0.001;0.002]	[-0.001;0.001]
Number of Teachers	-0.157	-0.120	0.583***	0.163	0.479***	0.212*	0.080	-0.259***
	[-0.174;0.195]	[-0.209;0.206]	[-0.202;0.217]	[-0.170;0.158]	[-0.179;0.177]	[-0.206;0.192]	[-0.193;0.209]	[-0.162;0.175]

Notes: Estimates use the main analysis sample. Each row represents the treatment-control mean differences in the value of a given variable. We subtract the overall average treatment effect for each variable taking the differences. Each column presents differences for the corresponding group of quartiles. Bootstrapped 90% confidence intervals in brackets. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table 5
Systematic Variation in Treatment Effects

	Base Model	Covariates of:			
		Students	Teachers	Schools	All
	(1)	(2)	(3)	(4)	(5)
Full-cost	1.396***	1.337***	1.325***	1.448***	1.395***
	(0.116)	(0.143)	(0.131)	(0.120)	(0.147)
Reduced-cost	0.738***	0.719***	0.791***	0.759***	0.896***
	(0.108)	(0.133)	(0.102)	(0.118)	(0.129)
Full-cost*1(BL Missing)		-0.248*			-0.279**
		(0.130)			(0.126)
Reduced-cost*1(BL Missing)		-0.148			-0.165
		(0.105)			(0.100)
Full-cost*1(Male)		-0.037			-0.032
		(0.103)			(0.102)
Reduced-cost*1(Male)		-0.102			-0.104
		(0.097)			(0.094)
Full-cost*Age		0.076			0.055
		(0.059)			(0.051)
Reduced-cost*Age		0.032			-0.013
		(0.058)			(0.051)
Full-cost*Baseline Test Score		0.085			0.077
		(0.135)			(0.132)
Reduced-cost*Baseline Test Score		0.069			0.071
		(0.145)			(0.141)
Full-cost*1(Male Teacher)			0.379		0.390
			(0.255)		(0.245)
Reduced-cost*1(Male Teacher)			-0.282		-0.277
			(0.247)		(0.240)
Full-cost*Teacher's Age			-0.014		-0.003
			(0.033)		(0.032)
Reduced-cost*Teacher's Age			0.000		-0.024
			(0.031)		(0.035)
Full-cost* Teacher's experience			-0.000		0.002
			(0.035)		(0.034)
Reduced-cost* Teacher's experience			0.005		0.040
			(0.030)		(0.036)
Full-cost*Years of Education			-0.038		-0.005
			(0.104)		(0.096)
Reduced-cost*Years of Education			-0.028		0.095
			(0.090)		(0.084)
Full-cost*School's Enrollment				0.001	0.001
				(0.001)	(0.001)
Reduced-cost*School's Enrollment				0.002	0.002
				(0.001)	(0.001)
Full-cost*Pupil-Teacher-Ratio				-0.002	0.001
				(0.018)	(0.017)
Reduced-cost*Pupil-Teacher-Ratio				-0.016	-0.012
				(0.013)	(0.013)
Full-cost*PLE Pass Rate				-0.800	-0.067
				(4.558)	(4.714)
Reduced-cost*PLE Pass Rate				-4.281	-4.552
				(4.801)	(5.201)
Full-cost*Number of Teachers				-0.043	-0.040
				(0.084)	(0.083)
Reduced-cost*Number of Teachers				-0.119*	-0.126*
				(0.071)	(0.070)
Observations	4,868	4,868	4,868	4,868	4,868
R-squared	0.179	0.181	0.186	0.193	0.202
Adj-R-Squared	0.170	0.171	0.175	0.183	0.188
Group*Year*Cohort FE	Yes	Yes	Yes	Yes	Yes

Notes: Estimates use the main analysis sample. Outcome is the Leblango reading score, standardized with respect to the control group. Regressions are estimated using equation 7. Each specification also includes main effects for all of the covariates that are interacted with the treatment indicators. All covariates interacted are de-meaned prior to estimation. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table 6
Post-LASSO Coefficients for Median Models Across 30 Sample Splits

	Full Covariate Set		Restricted Covariate Set	
	(1)	(2)	(3)	(4)
Full-cost	1.471***		1.749***	1.792***
	(0.134)		(0.263)	(0.264)
Reduced-cost	2.230**	0.631	0.866***	0.978***
	(0.976)	(1.080)	(0.101)	(0.183)
Full-cost*1(BL Missing)			-0.172*	-0.173*
			(0.096)	(0.094)
Reduced-cost*1(BL Missing)			-0.122*	
			(0.072)	
Full-cost*1(Male)			-0.029	-0.032
			(0.092)	(0.092)
Reduced-cost*1(Male)			-0.121	
			(0.076)	
Reduced-cost*Age				-0.018
				(0.040)
Full-cost*1(BL group 1)		-0.689***	-0.357	-0.394*
		(0.149)	(0.233)	(0.231)
Reduced-cost*1(BL group 1)				-0.135
				(0.138)
Full-cost*1(BL group 2)			-0.179	-0.174
			(0.324)	(0.325)
Full-cost*1(Male Teacher)	0.348**		0.648***	0.433*
	(0.157)		(0.214)	(0.240)
Reduced-cost*1(Male Teacher)				-0.289
				(0.224)
Full-cost*Teacher's Age	-0.030	-0.047		
	(0.027)	(0.042)		
Reduced-cost*Teacher's Age	-0.023	-0.000	-0.026	
	(0.025)	(0.022)	(0.031)	
Full-cost* Teacher's experience	0.013	0.001	-0.003	-0.002
	(0.028)	(0.045)	(0.016)	(0.014)
Reduced-cost* Teacher's experience			0.037	0.012
			(0.035)	(0.013)
Reduced-cost*Years of Education			0.126**	0.121**
			(0.056)	(0.052)
Full-cost*School's Enrollment			0.001**	0.001*
			(0.000)	(0.000)
Reduced-cost*Pupil-Teacher-Ratio		0.001	0.007	0.007
		(0.006)	(0.005)	(0.005)
Full-cost*PLE Pass Rate			-2.276	-0.548
			(5.020)	(3.833)
Reduced-cost*PLE Pass Rate		-9.370***	-9.262	-6.148
		(3.295)	(5.837)	(4.132)
Full-cost*Student's Age^2	-0.009	-0.065***		
	(0.016)	(0.019)		
Full-cost*ln(Student's Age)		0.618***		
		(0.205)		
Full-cost*ln(Teacher's Age)		0.704*		
		(0.386)		
Full-cost*Teacher's Experience^2	-0.000			
	(0.001)			
Full-cost*ln(Teacher's Experience)		-0.144		
		(0.387)		
Full-cost*1(Male Teacher)*School's Enrollment		-0.000		
		(0.001)		
Full-cost*1(Male Teacher)*Pupil-Teacher-Ratio	-0.028***			
	(0.010)			
Full-cost*1(Male Teacher)*Number of Teachers		0.040		
		(0.049)		
Full-cost*Teacher's Age*PLE Pass Rate	-0.201			
	(0.493)			

Table 6
Post-Lasso Coefficients Among Median Models in 30 Sample Splits (continued)

	Full Covariate Set		Restricted Covariate Set	
	(1)	(2)	(3)	(4)
Full-cost*Teacher's Age*Number of Teachers	0.009***	0.011***		
	(0.003)	(0.003)		
Full-cost*Teacher's Experience*Pupil-Teacher-Ratio	0.002***	-0.001		
	(0.001)	(0.001)		
Full-cost*Teacher's Experience*Number of Teachers		-0.008**		
		(0.004)		
Full-cost*Years of Education*School's Enrollment	-0.001***			
	(0.000)			
Full-cost*Years of Education*Number of Teachers	0.002	-0.036**		
	(0.019)	(0.018)		
Full-cost*School's Enrollment*PLE Pass Rate	-0.015	-0.009		
	(0.012)	(0.011)		
Full-cost*Pupil-Teacher-Ratio*PLE Pass Rate	0.614***	0.803***		
	(0.210)	(0.214)		
Reduced-cost*Student's Age^2		-0.039		
		(0.048)		
Reduced-cost*Student's Age^3		-0.001		
		(0.011)		
Reduced-cost*ln(Student's Age)	-0.128	0.335		
	(0.171)	(0.229)		
Reduced-cost*Teacher's Age^3		-0.000		
		(0.000)		
Reduced-cost*ln(Teacher's Age)	-0.321	-0.245		
	(0.338)	(0.382)		
Reduced-cost*Teacher's Experience^3	0.000***			
	(0.000)			
Reduced-cost*ln(Teacher's Experience)		0.300		
		(0.206)		
Reduced-cost*1(Male Teacher)*Pupil-Teacher-Ratio	-0.005			
	(0.006)			
Reduced-cost*Teacher's Age*Years of Education	-0.017***	-0.021**		
	(0.007)	(0.008)		
Reduced-cost*Teacher's Age*Pupil-Teacher-Ratio		-0.001		
		(0.001)		
Reduced-cost*Teacher's Experience*Pupil-Teacher-Ratio	0.001***			
	(0.000)			
Reduced-cost*Years of Education*Pupil-Teacher-Ratio	0.004*			
	(0.002)			
Reduced-cost*Years of Education*Number of Teachers		-0.007		
		(0.009)		
Reduced-cost*Pupil-Teacher-Ratio*PLE Pass Rate		0.014		
		(0.158)		
Reduced-cost*Pupil-Teacher-Ratio*Number of Teachers	0.002***			
	(0.001)			
Reduced-cost*PLE Pass Rate*Number of Teachers	-1.921***	-1.876***		
	(0.423)	(0.493)		
Observations	4,868	4,868	4,868	4,868
R-squared	0.226	0.227	0.197	0.197
Number of Selected Vars	23	28	15	15
Number of total Interactions	48	48	48	48
Percentage of Selected Variables	47.9%	58.3%	31.3%	31.3%
Adj-R-Squared	0.213	0.213	0.186	0.186
Mean full-cost CATE	1.433	1.521	1.385	1.397
Mean reduced-cost CATE	0.976	0.947	0.852	0.782
Median full-cost CATE	1.505	1.612	1.425	1.435
Median reduced-cost CATE	1.014	1.017	0.859	0.783
Group*Year*Cohort FE	Yes	Yes	Yes	Yes

Notes: Estimates use the main analysis sample. The table presents the coefficients selected by a cross-validated LASSO algorithm. Columns (1) and (2) present the full covariate set. Columns (3) and (4) present the restricted covariate set The presented sample splits for each covariate set are the ones with the median adjusted r-squared across the 30 sample splits. Standard Errors clustered at the school level and condition on stratification cells. All specifications include stratification cell fixed effects. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Table 7
Top 20 Variables Selected Most Often

Full Covariate Set		Restricted Covariate Set	
Variable (1)	Freq. (2)	Variable (3)	Freq (4)
Full-cost*Teacher's Age	27	Reduced-cost*PLE Pass Rate	28
Full-cost*Pupil-Teacher-Ratio*PLE Pass Rate	27	Full-cost*School's Enrollment	25
Full-cost*Teacher's Age*Number of Teachers	26	Full-cost*1(Male Teacher)	25
Full-cost*ln(Student's Age)	26	Full-cost*1(BL group 1)	23
Full-cost*Student's Age^2	25	Reduced-cost*1(Male Teacher)	23
Full-cost*ln(Teacher's Experience)	22	Full-cost*1(BL Missing)	20
Reduced-cost*ln(Student's Age)	21	Reduced-cost*Pupil-Teacher-Ratio	18
Full-cost*Teacher's Age^3	21	Reduced-cost* Teacher's experience	18
Full-cost*1(Male Teacher)*Pupil-Teacher-Ratio	20	Full-cost*Age	16
Full-cost*Years of Education*School's Enrollment	20	Reduced-cost*1(BL Missing)	13
Reduced-cost*ln(Teacher's Age)	19	Full-cost*Pupil-Teacher-Ratio	13
Reduced-cost*Teacher's Experience^3	18	Reduced-cost*Years of Education	13
Reduced-cost*Student's Age^2	18	Full-cost*PLE Pass Rate	10
Reduced-cost*Teacher's Age	17	Reduced-cost*1(Male)	9
Full-cost*ln(Teacher's Age)	17	Full-cost* Teacher's experience	9
Full-cost*Teacher's Experience^3	16	Full-cost*1(BL group 2)	8
Reduced-cost*PLE Pass Rate*Number of Teachers	16	Full-cost*Years of Education	7
Reduced-cost*Teacher's Age^3	15	Reduced-cost*School's Enrollment	7
Full-cost*Years of Education*Number of Teachers	15	Reduced-cost*Age	6
Full-cost*Teacher's Age*PLE Pass Rate	15	Full-cost*1(Male)	6

Notes: This table displays the 20 variables included most often across the 30 sample splits, and the number of times they are included. Columns (1) and (2) refer to the full covariate set, whereas columns (3) and (4) refer to the restricted covariate set.

Table 8
Characteristics by Sign of ACATE

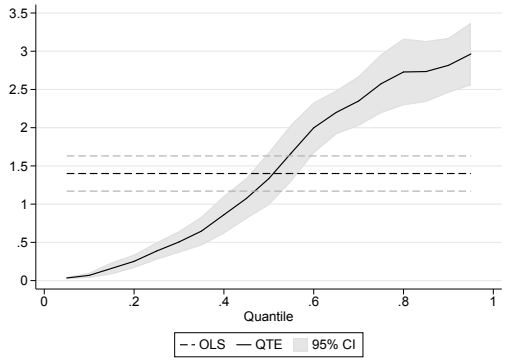
		Full-Cost			Reduced-Cost		
		ACate < 0 (1)	ACate ≥ 0 (2)	Diff. (3)	ACate < 0 (4)	ACate ≥ 0 (5)	Diff. (6)
Students	Stand. Baseline result	-0.14	-0.09	-0.05***	-0.10	-0.09	-0.01***
	1(BL Missing)	0.57	0.51	0.07	0.63	0.50	0.13
	1(Male)	0.49	0.50	-0.02***	0.51	0.50	0.00
	Age	9.24	9.83	-0.58***	9.30	9.83	-0.53***
Teachers	1(Male Teacher)	0.45	0.64	-0.19***	0.53	0.64	-0.11***
	Teacher's Age	43.82	39.35	4.47	46.86	39.20	7.66
	Teacher's Experience	16.57	12.83	3.75	15.62	12.82	2.80
	1(Teacher's Experience <=4)	0.12	0.17	-0.05***	0.07	0.17	-0.10***
	1(Teacher's Experience >4)	0.88	0.83	0.05	0.93	0.83	0.10
	Years of Education	16.01	14.62	1.39	15.65	14.62	1.03
Schools	School's Enrollment	1208.61	891.17	317.45	1244.79	886.41	358.38
	Pupil-Teacher-Ratio	73.52	69.08	4.44	55.22	69.69	-14.47***
	PLE Pass Rate	0.46	0.32	0.14	0.52	0.32	0.20
	Number of teachers	19.10	13.66	5.44	24.59	13.40	11.19

Notes: Estimates use the main analysis sample. ACATEs are the average predicted treatment effect across all 30 sample splits for the full covariate set. Columns (1) to (3) refer to the full-cost version of the program. Columns (4) to (6) refer to the reduced-cost version. The presented sample splits for each model are the ones with the median adjusted r-squared across the 30 sample splits. All specifications include stratification cell fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses: $*p < 0.1; **p < 0.05; ***p < 0.01$.

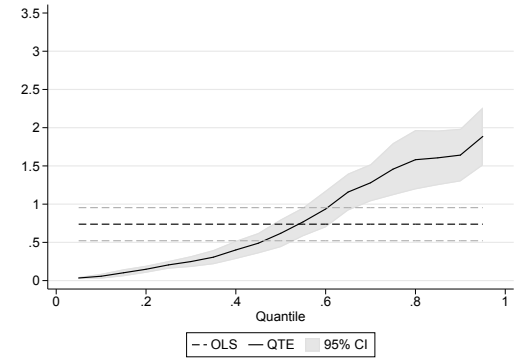
Figures

Figure 1
Quantile Treatment Effects

(a) Full-cost Program

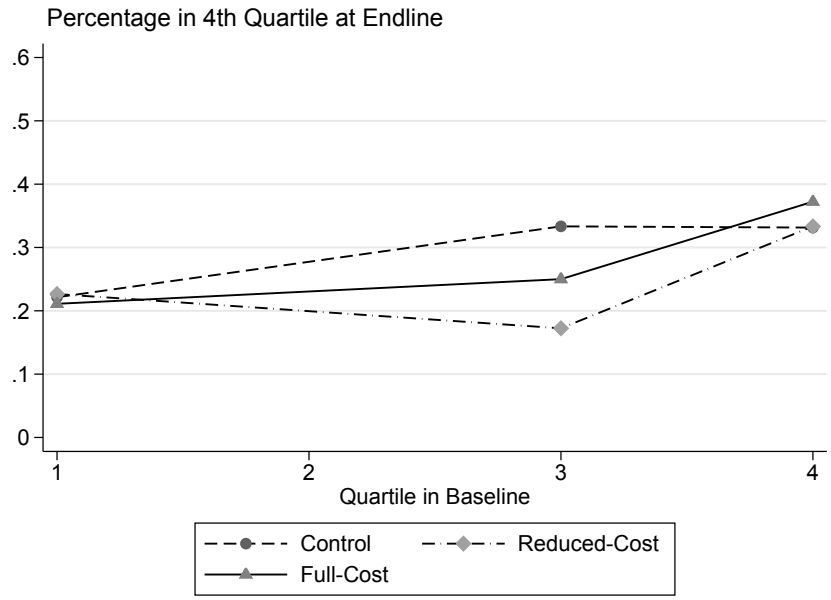


(b) Reduced-cost Program



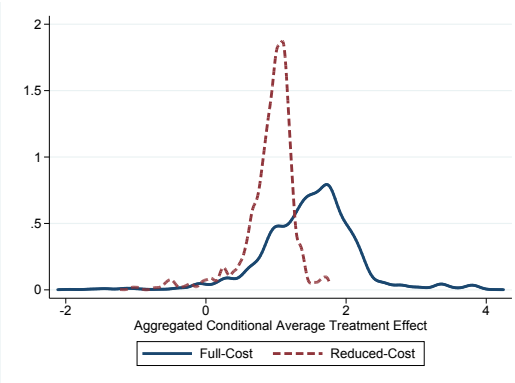
Notes: Estimates use the main analysis sample. Outcome is the Leblango reading test score index standardized with respect to the control group. Solid lines are quantile treatment effect estimates as described in Section 6.2.2; gray regions are bootstrapped 95% confidence intervals. The dark dashed line is the average treatment effect, with the 95% confidence interval indicated via light dashed lines.

Figure 2
Test Score Transition



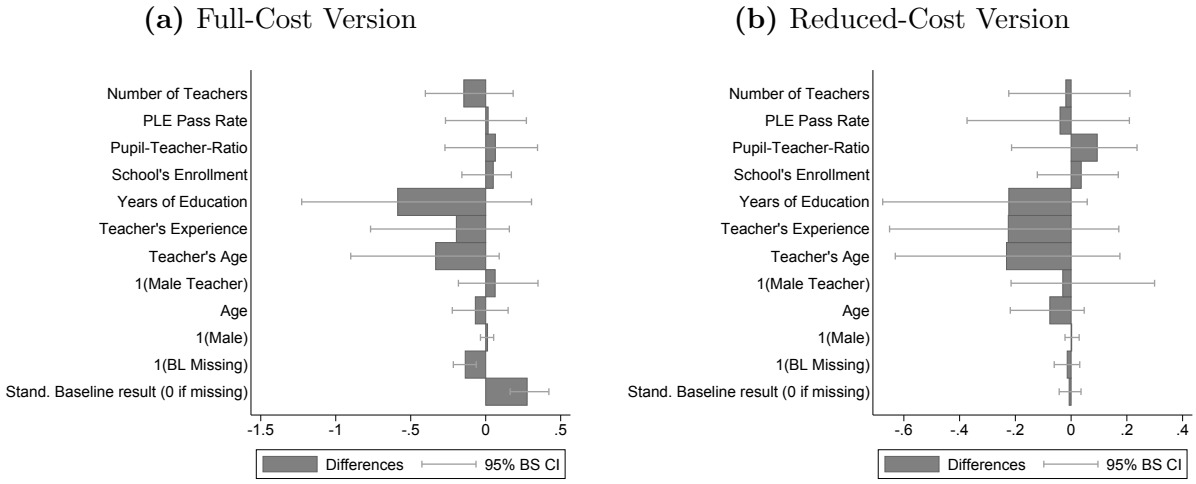
Notes: Sample is 2,395 students with a complete set of moderators and baseline results (723 control, 849 full-cost treatment, 823 reduced-cost treatment). Graph presents the share of students from each quartile of baseline scores who end up in the fourth quartile of endline scores.

Figure 3
Kernel Density of ACATEs



Notes: Estimates use the main analysis sample. Kernels estimated using optimal bandwidths that minimize the mean squared error, which equal 0.0886 for the full-cost and 0.0392 for the reduced-cost version. ACATEs are the average predicted treatment effect across all 30 sample splits for the full covariate set.

Figure 4
Mean Differences in CATEs by Student Characteristics

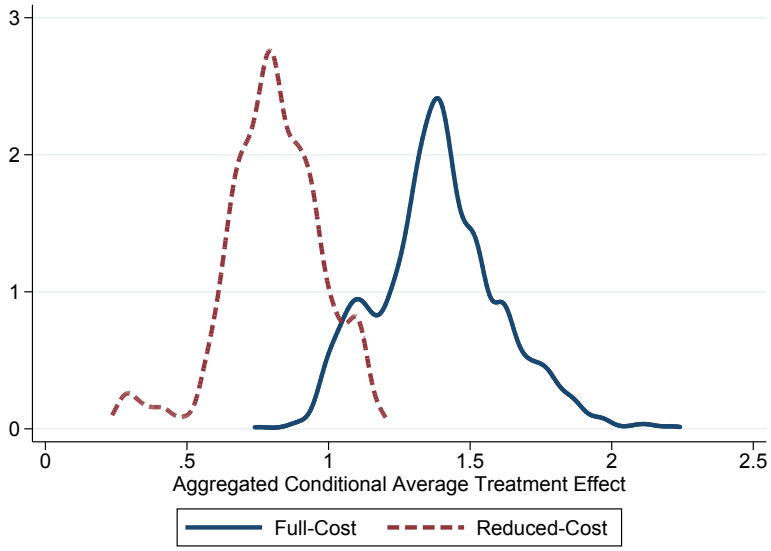


Notes: Estimates use the main analysis sample. The graph presents the average difference of the ACATEs, for the full covariate set, by whether various characteristics are high or low. "High" means the variable is equal to one for binary characteristics, or is above the median for continuous characteristics; "low" is the complement of high. Whiskers indicate 95% confidence intervals for the difference between high and low, computed using 200 bootstrap replications. Differences are computed using the ACATEs estimated with the full covariate set.

A Online Appendix

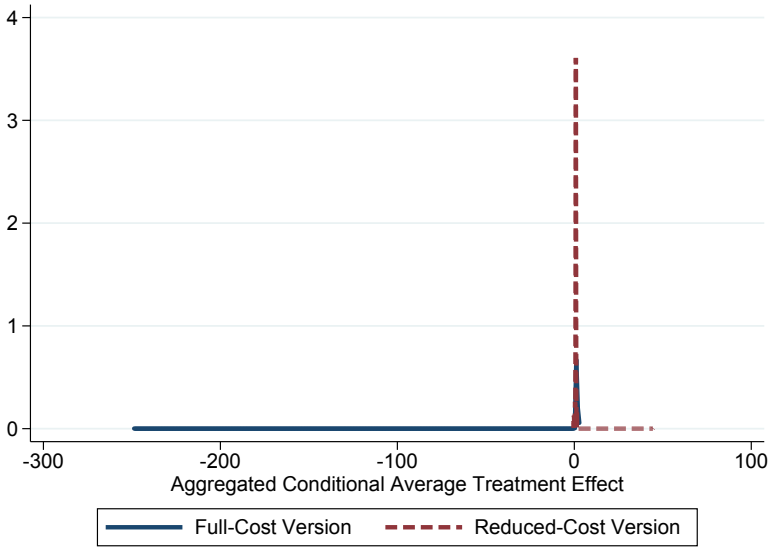
A.1 Online Appendix Figures

Online Appendix Figure 1
Kernel Density of ACATEs for Restricted Model



Notes: Estimates use the main analysis sample. Kernels estimated using optimal bandwidths, which correspond to 0.0321 for the full-cost and 0.0259 for the reduced-cost version. The presented sample splits for each model are the ones with the median adjusted r-squared across the 30 sample splits. All specifications include stratification cell fixed effects.

Online Appendix Figure 2
Distribution of ACATEs without Discretizing Baseline Exam Scores



Notes: Estimates use the main analysis sample. ACATEs for the full covariate set. Kernels estimates use optimal bandwidths, which correspond to 0.0547 for the full-cost and 0.0194 for the reduced-cost version.

A.2 Online Appendix Tables

Online Appendix Table 1
Sample Composition

	Control	Reduced-Cost Version	Full-Cost Version	Total
	(1)	(2)	(3)	(4)
Initial Sample	1,974	2,020	2,112	6,106
Missing Endline Scores	1	0	0	1
Missing Moderators	546	339	352	1,237
Main Analysis Sample	1,427	1,681	1,760	4,868

Online Appendix Table 2
Balance Across Treatment Arms

	Control	Reduced-Cost Version	Full-Cost Version	p-values		
				(1)-(3)	(1)-(2)	Joint-F
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline Score Index	-0.11	-0.10	-0.08	0.193	0.648	0.410
1(BL Missing)	0.49	0.51	0.52	0.136	0.222	0.270
1(Male)	0.51	0.49	0.51	0.689	0.180	0.406
Age	9.78	9.83	9.82	0.506	0.323	0.606
1(Male Teacher)	0.55	0.61	0.73	0.054	0.501	0.118
Teacher's Age	40.72	39.98	37.95	0.069	0.619	0.166
Teacher's Experience	14.54	13.14	11.39	0.012	0.267	0.042
Years of Education	14.56	14.50	14.88	0.090	0.760	0.101
Overall Pupil Enrollment	875.66	959.88	859.58	0.698	0.043	0.078
Pupil-Teacher-Ratio	67.35	72.86	67.16	0.951	0.069	0.144
PLE Pass Rate	0.05	0.05	0.05	0.642	0.950	0.826
Overall number of teachers	13.97	13.93	13.50	0.419	0.939	0.671

Notes: Estimates use the main analysis sample. Columns (1)-(3) present means for the control, reduced-cost, and full-cost versions, respectively. Column (4) presents p-values for the null of equality of means between the control and the full cost version, column (5) presents p-values for the null of equality of means between the control and the reduced-cost version of the program, and column (6) presents the p-value for a joint F-test of the equality of means across all three study arms. The p-values are obtained running a regression of the given outcome on full- and reduced cost-version dummies including stratification fixed effects, and clustering at the school level. The baseline score index corresponds to a standardized measure where the missing have been replaced by zeroes.

Online Appendix Table 3

Performance of Bootstrap Standard Errors for the Estimated Standard Deviation of Impacts

Parameter	(1) Sample 1 1,760 Treatments 1,427 Controls	(2) Sample 2 1,681 Treatments 1,427 Controls
Population impact standard deviation	0.000	0.000
Mean of data sample estimates	0.064	0.066
Std. Dev. of data sample estimates	0.025	0.026
Mean of boostrap standard errors	0.010	0.011
Std. dev. of bootstrap standard errors	0.008	0.009

Notes: Estimates are based on 250 data samples of the indicated size. All the sample are drawn from the control group. Faux treatment groups are created by adding 0.5 to an independent draw from the control group. Percentiles of the treatment and control groups samples are then calculated. Estimates of impact standard deviation are equivalent to the standard deviation of the difference between the percentiles of the control and the faux treatment group. Rows two and three present the mean and standard deviation, respectively, of such estimates across the 250 data samples. Bootstrap standard errors are calculated by drawing 250 bootstrap samples for each data sample. Rows three and four present the mean and standard deviation of the bootstrap standard errors across the 250 data samples.

Online Appendix Table 4

Monte Carlo p-values for the Null Hypothesis of a Zero Impact Standard Deviation

P-Value	(1) Cutoff Value for 1,760 Treatments 1,427 Controls	(2) Cutoff Value for 1,681 Treatments 1,427 Controls
0.500	0.056	0.057
0.400	0.062	0.062
0.300	0.068	0.069
0.200	0.076	0.077
0.100	0.088	0.089
0.050	0.099	0.101
0.010	0.123	0.126
0.001	0.157	0.153
0.000	0.193	0.193

Notes: Estimates are based on 10,000 random samples of the indicated size. All the sample are drawn from the control group.

Online Appendix Table 5
Test of Equality of Quantile Treatment Effects

Statistics	Full-cost Program (1)	Reduced-cost Program (2)
F-Stat	1,091.873	59.977
	(0.000)	(0.000)

Notes: This table presents the F-statistic and the associated p-value (in parentheses) for an equality of coefficients test. The tested coefficients correspond to those of the 19 ventiles.

Online Appendix Table 6

Unconditional Treatment-Control Differences at Quantiles of the Outcome Distribution

	Full-Cost				Reduced-Cost			
	0-25th Perc.	25-50th Perc.	50-75th Perc.	75-100th Perc.	0-25th Perc.	25-50th Perc.	50-75th Perc.	75-100th Perc.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline Test Score	-0.035*	-0.003	0.002	0.035	-0.016	0.005	-0.012	0.023
	[-0.029;0.032]	[-0.032;0.035]	[-0.042;0.042]	[-0.081;0.078]	[-0.029;0.031]	[-0.033;0.033]	[-0.042;0.041]	[-0.077;0.075]
1(BL Missing)	0.068*	-0.084**	0.030	-0.015	0.046	-0.072**	0.029	-0.006
	[-0.059;0.057]	[-0.060;0.060]	[-0.060;0.061]	[-0.059;0.062]	[-0.057;0.056]	[-0.059;0.061]	[-0.058;0.061]	[-0.057;0.059]
1(Student Male)	0.047	-0.057	-0.036	0.043	0.081**	-0.042	-0.072**	0.028
	[-0.065;0.058]	[-0.064;0.059]	[-0.060;0.057]	[-0.057;0.059]	[-0.057;0.061]	[-0.062;0.059]	[-0.062;0.056]	[-0.055;0.058]
Student's Age	-0.028	-0.099	0.159*	-0.036	-0.021	-0.075	0.071	0.022
	[-0.148;0.145]	[-0.145;0.140]	[-0.150;0.129]	[-0.126;0.122]	[-0.146;0.136]	[-0.141;0.128]	[-0.148;0.136]	[-0.127;0.127]
1(Male Teacher)	-0.097***	-0.026	0.020	0.106***	-0.002	0.014	-0.039	0.029
	[-0.058;0.056]	[-0.055;0.062]	[-0.060;0.061]	[-0.054;0.057]	[-0.054;0.057]	[-0.053;0.056]	[-0.054;0.058]	[-0.052;0.054]
Teacher's Age	-0.555	0.930	-0.516	0.199	-1.385**	0.942	-0.038	0.564
	[-1.010;1.033]	[-0.991;1.048]	[-0.940;0.984]	[-0.873;0.863]	[-0.991;0.969]	[-0.993;0.979]	[-0.995;0.966]	[-0.846;0.857]
Teacher's Experience	-0.050	0.657	-0.587	0.001	-0.617	0.754	-0.215	0.117
	[-0.865;0.828]	[-0.840;0.891]	[-0.812;0.884]	[-0.847;0.802]	[-0.851;0.845]	[-0.835;0.840]	[-0.914;0.845]	[-0.796;0.807]
Years of Education	-0.040	0.135	-0.061	-0.034	-0.019	0.106	0.041	-0.126
	[-0.148;0.141]	[-0.156;0.150]	[-0.151;0.137]	[-0.146;0.135]	[-0.149;0.140]	[-0.148;0.147]	[-0.147;0.146]	[-0.140;0.136]
School's Enrollment	44.819**	-0.629	-27.224	-17.755	27.210	40.559*	5.107	-73.242***
	[-34.890;34.044]	[-36.384;37.747]	[-36.374;36.131]	[-38.372;33.186]	[-35.002;34.717]	[-37.753;38.134]	[-39.658;36.627]	[-36.990;32.410]
Pupil-Teacher-Ratio	-0.616	0.527	-1.785	1.997*	-2.097	2.162	-1.013	1.112
	[-2.410;2.117]	[-2.236;2.347]	[-2.220;2.275]	[-1.962;1.915]	[-2.233;2.108]	[-2.206;2.202]	[-2.406;2.317]	[-1.955;1.911]
PLE Pass Rate	-0.001	0.003	0.001	-0.003*	0.005***	0.002	-0.001	-0.007***
	[-0.003;0.003]	[-0.003;0.003]	[-0.003;0.003]	[-0.003;0.003]	[-0.003;0.003]	[-0.003;0.003]	[-0.003;0.003]	[-0.003;0.003]
Number of Teachers	0.564	0.170	-0.176	-0.568	0.616	0.516	0.177	-1.321***
	[-0.690;0.715]	[-0.724;0.683]	[-0.691;0.616]	[-0.677;0.632]	[-0.682;0.695]	[-0.667;0.657]	[-0.710;0.681]	[-0.669;0.657]

Notes: Estimates use the main analysis sample. Each row represents the unconditional treatment-control mean differences in the value of a given variable. We subtract the overall average treatment effect for each variable taking the differences. Each column presents differences for the corresponding group of quartiles. Bootstrapped 90% confidence intervals in brackets: $*p < 0.1; **p < 0.05; ***p < 0.01$.

Online Appendix Table 7

Results of Conducting Principal Component Analysis on Moderators

	Eigenvectors											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Eigenvalue	2.13	1.81	1.36	1.24	1.12	1.03	0.93	0.85	0.71	0.68	0.11	0.03
Prop. of Var. Explained	0.18	0.15	0.11	0.10	0.09	0.09	0.08	0.07	0.06	0.06	0.01	0.00
Cum. Prop. Explained	0.18	0.33	0.44	0.54	0.64	0.72	0.80	0.87	0.93	0.99	1.00	1.00
Baseline Test Score	-0.01	0.03	-0.29	0.65	-0.03	0.04	-0.17	-0.07	0.67	-0.08	-0.00	-0.00
1(BL Missing)	0.04	-0.15	0.35	-0.56	0.17	0.03	-0.11	0.15	0.68	-0.10	-0.00	0.01
1(Student Male)	-0.06	0.00	0.02	0.05	0.47	0.59	0.51	-0.37	0.03	-0.11	0.00	0.00
Student's Age	-0.10	0.05	0.15	0.24	0.57	0.23	-0.37	0.59	-0.18	0.12	0.01	0.01
1(Male Teacher)	0.13	0.36	0.11	-0.10	-0.25	0.46	-0.31	-0.23	0.08	0.64	-0.01	0.00
Teacher's Age	0.55	0.35	0.03	0.01	0.17	-0.15	0.03	-0.02	0.00	-0.12	0.71	-0.00
Teacher's Experience	0.54	0.34	-0.01	0.01	0.22	-0.21	0.14	0.04	0.03	0.00	-0.69	-0.00
Years of Education	0.27	0.03	-0.07	-0.04	-0.44	0.55	-0.05	0.35	-0.08	-0.53	-0.08	-0.02
School's Enrollment	0.29	-0.41	0.47	0.31	-0.14	0.01	0.18	0.05	-0.01	0.20	0.01	-0.58
Pupil-Teacher-Ratio	-0.20	0.30	0.50	0.24	-0.25	-0.07	0.44	0.27	0.10	0.07	0.04	0.46
PLE Pass Rate	0.09	-0.13	-0.53	-0.16	-0.04	0.06	0.45	0.48	0.15	0.44	0.11	-0.04
Number of Teachers	0.40	-0.57	0.03	0.09	0.03	0.06	-0.11	-0.11	-0.07	0.13	-0.01	0.67

Notes: Estimates use the main analysis sample. Columns represent the 12 estimated principal components for the 12 moderators we use to examine systematic variation in treatment effects.

Online Appendix Table 8
Fréchet-Höffding Bounds
Removing Systematic Variation

	Full-cost Program		Reduced-cost Program	
	Rank Preservation (1)	Rank Inversion (2)	Rank Preservation (3)	Rank Inversion (4)
Percentiles under control status				
5th	-0.238	5.578	-0.124	4.623
	(0.043)	(0.116)	(0.051)	(0.145)
25th	0.401	3.222	0.295	2.226
	(0.072)	(0.158)	(0.048)	(0.122)
50th	1.382	1.382	0.770	0.770
	(0.175)	(0.175)	(0.084)	(0.084)
75th	2.481	-0.340	1.485	-0.446
	(0.180)	(0.118)	(0.153)	(0.102)
95th	2.911	-2.905	1.957	-2.791
	(0.203)	(0.169)	(0.208)	(0.170)
Impact Standard Deviation	1.079	2.640	0.660	2.257
	(0.023)	(0.014)	(0.018)	(0.010)
Outcome Correlation	0.925	-0.689	0.971	-0.619
	(0.014)	(0.020)	(0.008)	(0.015)
Fraction Positive	0.889	0.697	0.909	0.677
	(0.021)	(0.017)	(0.021)	(0.018)

Notes: Estimates use the main analysis sample. Table presents Fréchet-Höffding bounds of the EGRA Leblango test score index distribution, computed as described in Section 5.1. Prior to estimating the bounds, we subtract each student's estimated systematic variation from the outcome. Standard errors computed using 1,000 bootstrap replications.

Online Appendix Table 9
Values for the Penalty Coefficient (λ) Across Sample Splits

Sample Split	Full Covariate Set (1)	Restricted Covariates Set (2)
1	147.422	5.345
2	163.343	5.692
3	164.344	2.741
4	167.077	2.250
5	196.729	5.084
6	80.171	7.475
7	150.563	1.842
8	51.673	4.322
9	192.289	9.323
10	41.576	5.125
11	499.145	2.822
12	156.058	11.913
13	65.231	6.491
14	105.815	6.182
15	240.514	9.094
16	100.837	8.116
17	136.633	8.957
18	74.747	8.763
19	274.670	3.692
20	637.811	5.862
21	173.471	5.396
22	86.077	3.698
23	100.251	2.282
24	122.340	10.840
25	95.264	2.021
26	138.474	13.838
27	231.067	19.288
28	104.896	3.779
29	84.036	0.293
30	41.910	13.236

Note: The penalty parameters correspond to those that minimize the mean-squared prediction error after a 10 K-fold cross-validation procedure.

Online Appendix Table 10
Descriptive Statistics of Aggregated CATEs

	Outcome	Mean (1)	Median (2)	S.D. (3)	Min. (4)	Max. (5)	Mean S.E. (6)
Full Covariate Set	Full-Cost Model	1.489	1.530	0.682	-2.115	4.247	0.522
	Reduced-Cost Model	0.911	0.990	0.373	-1.224	1.757	0.356
Restricted Covariate Set	Full-Cost Model	1.389	1.384	0.217	0.739	2.242	0.306
	Reduced-Cost Model	0.806	0.806	0.172	0.236	1.202	0.300

Notes: Columns (1) to (5) present the mean, median, standard deviation, minimum, and maximum, respectively. Column (6) presents the standard error of the mean computed using 200 bootstrap replications. For each bootstrap sample we average across the ACATEs all 30 sample splits, and then take the standard deviation of the resulting 200 averages.

Online Appendix Table 11
Fréchet-Höffding Bounds
Removing Systematic Variation Using the LASSO Full Model

	Full-cost Program		Reduced-cost Program	
Percentiles under control status	Rank Preservation (1)	Rank Inversion (2)	Rank Preservation (3)	Rank Inversion (4)
5th	3.135	8.852	-0.015	4.566
	(0.056)	(0.099)	(0.057)	(0.115)
25th	3.865	6.585	0.299	2.287
	(0.086)	(0.115)	(0.050)	(0.116)
50th	4.790	4.790	0.825	0.825
	(0.123)	(0.123)	(0.091)	(0.091)
75th	5.844	3.124	1.546	-0.442
	(0.147)	(0.126)	(0.152)	(0.110)
95th	6.185	0.468	1.899	-2.682
	(0.198)	(0.177)	(0.193)	(0.177)
Fraction Positive	1.000	0.970	0.949	0.677
	(0.000)	(0.007)	(0.017)	(0.018)
Impact Standard Deviation	1.021	2.589	0.648	2.239
	(0.023)	(0.013)	(0.019)	(0.011)
Outcome Correlation	0.923	-0.708	0.968	-0.626
	(0.013)	(0.020)	(0.009)	(0.018)

Notes: Estimates use the main analysis sample. Table presents Fréchet-Höffding bounds of the EGRA Leblango test score index distribution, computed as described in Section 5.1. Prior to estimating the bounds, we subtract each student's average ACATE as computed in the LASSO full model from her respective outcome. Standard errors computed using 1,000 bootstrap replications.