ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Some children left behind: Variation in the effects of an educational intervention ☆

Julie Buhl-Wiggers [a], Jason T. Kerwin [b], Juan Muñoz-Morales [c], Jeffrey Smith [d,*], Rebecca Thornton [e]

[a] Department of Economics, Copenhagen Business School, Denmark
[b] Department of Applied Economics, University of Minnesota and J-PAL, United States of America
[c] IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Économie Management, F-59000 Lille, France
[d] Department of Economics, University of Wisconsin, United States of America
[e] Department of Economics, University of Illinois, United States of America

ARTICLE INFO

Article history:
Received 13 August 2020
Received in revised form 30 October 2021
Accepted 21 December 2021
Available online xxxx

JEL classification:
Codes
C18
C21
I21
I25
J24

Keywords:
Treatment effect heterogeneity
Machine learning
Education programs

ABSTRACT

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.

1. Introduction

This paper examines treatment effect heterogeneity in the context of an educational intervention implemented in northern Uganda. Kerwin and Thornton (2021) and Buhl-Wiggers et al. (2018a) show that the intervention – the Northern

☆ We thank participants at the Heckman 75th Birthday Conference, seminar audiences at Aarhus, CESifo, Copenhagen Business School, and RISE, as well as Natalie Bau, Jishnu Das, Paul Glewwe, Lois Miller, Paul Niehaus, Lant Pritchett and three anonymous referees for helpful comments, Brigham Frandsen for assistance in implementing the Frandsen–Lefgren bounds, and Joseph Cummins for sharing his rank similarity test code. 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. We are grateful for funding from DFID/ESRC Raising Learning Outcomes Grant ES/M004996/2, Wellspring, and the International Growth Centre. The original data collection for this project is registered with the AEA RCT Registry under registration number AEARCTR-0000021. The data and code for this paper are available on the Harvard Dataverse as Buhl-Wiggers et al. (2022). The usual disclaimer applies.

* Corresponding author.
E-mail addresses: jubu.eco@cbs.dk (J. Buhl-Wiggers), jkerwin@umn.edu (J.T. Kerwin), j.munoz@ieseg.fr (J. Muñoz-Morales), econjeff@ssc.wisc.edu (J. Smith), rebeccat@illinois.edu (R. Thornton).

https://doi.org/10.1016/j.jeconom.2021.12.010
0304-4076/

Uganda Literacy Project (NULP) – has an extraordinarily large average treatment effect, especially relative to other education interventions (McEwan, 2015; Evans and Yuan, 2019). At the same time, 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. Understanding treatment effect heterogeneity can shed important light on how interventions work, for whom they work, and how they affect inequality.

There is broad concern about some students being "left behind" in learning (Rudalevige, 2003). This issue was highlighted in the United States by the No Child Left Behind Act of 2002, but also looms large in developing countries. A recent World Development Report focused solely on the "learning crisis" (World Bank, 2018): while enrollment rates are high, many students learn almost nothing in school (Boone et al., 2014; Piper, 2010). Similarly, Sustainable Development Goal 4 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 using classical statistical bounds due to Fréchet (1951) and Höffding (1940). These "FH bounds" allow us to bound the variance of the treatment effects; related formal statistical tests reject the null of a common treatment effect. Second, we consider what we can learn about treatment effect heterogeneity by imposing two additional assumptions. One assumption, "mutual stochastic increasingness" of the joint distribution of treated and untreated outcomes, allows us to bound the average treatment effects at particular quantiles of the outcome distributions [hereinafter "FL bounds" after Frandsen and Lefgren (2021)]. The other assumption, rank preservation, allows us to analyze treatment effects on particular quantiles under the *status quo* (i.e. for the control group).[1] Third, we look for moderators that capture meaningful variation in the treatment effect, 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 Chernozhukov et al. (2020).

Our core finding is that the effects of the program vary widely across individual students. The estimated lower bound on the standard deviation of treatment effects exceeds one standard deviation. Despite a massive average gain of 1.4 SDs, if the treatment effects are normally distributed then the intervention harms over 10 percent of students, while 29 percent of students experience individual gains in excess of 2.0 SDs. These results imply that the difference between the 95th and 5th percentile of treatment effects for students within the NULP exceeds the difference in the average effects between the least- and most-beneficial interventions reviewed in McEwan (2015). We also compare the variation in the treatment effects of the NULP to the change in average effects when the program is modified to lower costs by removing non-essential inputs and doing the teacher training more cheaply. The average effect of the reduced-cost version of the NULP equals 0.74 SDs. Thus the variation in treatment effects within the original version of the NULP is over four times as large as the gap between the two versions of the program.

At the same time, we make little headway in systematizing the treatment effect heterogeneity the data clearly contain. The FL bounds suggest that negative average treatment effects – to the extent they exist at all – occur at the top of the outcome distribution. Rank preservation provides 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 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 suggest that the extensive literature documenting the average effects of education interventions is fundamentally insufficient, generating very little information about how the effects of individual interventions vary across students. Eight recent reviews of "what works" in education in developing countries collectively cover hundreds of randomized trials in dozens of countries; most individual studies and these reviews focus almost entirely on average treatment effects.[2] Even re-analyses of the raw data may yield limited evidence, since studies are commonly powered to detect only average effects (Glewwe and Muralidharan, 2016). Examples of studies that do examine heterogeneity include Jackson and Makarin (2018), who use a conditional quantile treatment regression approach to show that the lesson plans matter more for weaker teachers, Glewwe et al. (2009), who find that textbooks only improve scores for the strongest students, and Moshoeshoe (2015) who investigates heterogeneity in the effects of class size reductions in Lesotho.

This 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 et al. [hereinafter "HSC"] (1997), Djebbari and Smith (2008), and Bitler et al. (2017). Not only does northern Uganda differ greatly from the United States and 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 et al. (2017). Our findings regarding the clear presence of "essential

[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 et al. (2008).

[2] The eight reviews are: Glewwe et al. (2013), Kremer et al. (2013), Krishnaratne et al. (2013), Ganimian and Murnane (2014), McEwan (2015), Evans and Popova (2016), Glewwe and Muralidharan (2016) and Conn (2017). Four of these reviews – Ganimian and Murnane (2014), Evans and Popova (2016), Glewwe and Muralidharan (2016) and Conn (2017) – discuss systematic heterogeneity in the effects of one specific intervention, although each chooses to highlight a different intervention for this purpose. 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. Evans and Yuan (2018) review 281 evaluations with learning outcomes, conducted between 2000 and 2016; 33 percent presented results separately by gender, 23 differentiated effects by baseline achievement, and only 11 percent differentiated effects by socio-economic status.

heterogeneity" (Heckman et al., 2006), combined with our general failure to systematize that heterogeneity via observed moderators (even with machine learning methods) defines an agenda for future evaluations of educational interventions: empiricists should collect improved candidate moderators and applied theorists should devote themselves to motivating new moderators.[3]

Methodologically, our paper represents only the second empirical application of FL (2021) bounds and arguably the first with a data set of meaningful size. While numerous recent papers examine treatment effect heterogeneity using one of the vast array of competing machine learning algorithms currently in circulation, we add value by comparing traditional a priori methods to machine learning algorithms. We do this within the context of a broader discussion of theories of treatment effect moderation. 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.

The remainder of the paper takes a familiar course. We describe the NULP intervention in Section 2 and describe the data we analyze in Section 3; Section 4 presents the average treatment effects of the program for reference. Section 5 estimates 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. Educational context in Uganda

Our study takes place in the Lango sub-region of northern Uganda, one of the poorest regions of the country. The primary education system – running from P1 (first grade) to P7 (seventh grade) – in northern Uganda faces major challenges. The pupil-to-teacher ratio is about 58:1 and absenteeism is high. On an average day about 28 percent of teachers and 24 percent of students miss school (Bold et al., 2017; Uwezo, 2016). The majority of schools lack electricity, though nearly all have a latrine. The central government provides an annual allocation to each school for teaching and learning materials, extra-curricular activities, school management, and school administration. Still, these funds, combined with "contributions" from parents, continue to leave many schools in dire need.

Until fairly recently, teaching in Ugandan schools reflected Uganda's British colonial past—often entirely in English with a call-and-response pedagogy. In 2007, the government of Uganda implemented a new primary education curriculum aimed at improving on this history. The new curriculum includes two important features. First, students in P1–P3 should be taught in the main mother tongue of their area with a transitional year in P4 leading to full English instruction starting in P5. Second, teachers should devote an hour to literacy lessons each day, with the first half hour on reading and the second on writing. In practice, many teachers have 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 et al. (2016).

Poor schools, combined with a history of civil conflict, lead to poor outcomes: the adult literacy rate in the Lango sub-region sits just above 71 percent (Uganda Bureau of Statistics, 2017). Piper (2010) found that 80 percent of students in the sub-region could not read a single word of Leblango (the local language) at the end of P2 and 50 percent could not at the end of P3.

2.2. The NULP intervention

From 2009 to 2013, Mango Tree – a private, for-profit, educational tools company – developed a program, called the Northern Uganda Literacy Project (NULP), focused on mother tongue literacy in P1 to P3. The program consisted of four main features. First, it provided teachers with intensive training in teaching mother tongue literacy including residential training sessions as well as in-class coaching visits.[4] Second, the NULP provided classroom materials including primers (textbooks that follow the curriculum), readers (books for reading practice), teacher guides with scripted daily lesson plans, chalk slates for writing practice, 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

[3] Predicting treatment effects will likely mean going beyond the set of potential moderators typically available to schools or educational authorities in their administrative data. If researchers collect better moderators, they could use them to alter the design of programs to trim the lower tail of treatment effects while holding steady, or even increasing, the average gains.

[4] Over the school year there were 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 with previous experience teaching the NULP instruction model, or coordinating center tutors employed by the Ministry of Education.

was introduced as a subject in P1; written English was 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. Finally, the NULP model engaged with the surrounding community to promote the benefits of mother-tongue instruction using a radio program, and held school meetings to train parents on how to support their children's learning at home. According to Kerwin and Thornton (2021), 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.

2.3. Bundled interventions

The NULP bundles several complementary interventions, which we study as a combined whole. This is the most policy-relevant way to examine the program: Delavallade et al. (2021) point out that while the majority of programs actually *implemented* in developing countries involve a packaged bundle of education inputs, most evaluations study the effectiveness of a single intervention.[5] Such 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). For analyses of which parts of the NULP program matter most, see Kerwin and Thornton (2021) on potential complementarities between the program components and Buhl-Wiggers et al. (2018a) for an examination of what happens when certain program components are removed.

3. Evaluation

3.1. Evaluation design

The NULP was evaluated over four academic years running from 2013 through 2016; 38 schools were selected to be part of the study 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.[6] Randomization took place within pre-defined strata of three schools.[7] The reduced-cost version embodied two main changes. First, 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. Second, teachers received fewer support visits throughout the year.[8] 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.[9]

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

3.2. Analytical sample

In this paper, we focus solely on students who entered P1 in 2014 in one of the 128 study schools. Because the intervention was rolled out to grades P1, P2 and P3 across years, students in treatment schools were exposed to three full academic years of either the full- or reduced-cost NULP. Focusing on just one cohort of students avoids mechanical variation in treatment intensity resulting from differing amounts of exposure to the program.

In 2014, 100 P1 students were sampled from each school, stratified by sex and classroom. Students were either sampled at the beginning or end of the school year (we call the latter "top-up students").[10]

Our "main analysis sample" involves two additional restrictions. First, we require that students have valid test scores at the end of P3 in 2016. Second, to analyze moderators, we require complete data on all of the variables we use as

[5] Exceptions include the Primary Math and Reading (PRIMR) Initiative in Kenya (Piper et al., 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 et al., 2009) and teacher training (Cilliers et al., 2020).

[6] In 2013, eligibility required that a school has two P1 classrooms, lockable classrooms, a head teacher regarded as "engaged", less than 135 students/teacher, and be located less than 20 km from the coordinating center for the school. In 2014, the only requirements for participation in the study were to have less than 150 students/teacher and be located at most 22 km from the coordinating center.

[7] Stratification groups (i.e. strata) were defined based on P1 enrollment, coordinating center, and distance to coordinating center headquarters.

[8] Buhl-Wiggers et al. (2018a) evaluate the scale-up of the NULP program.

[9] Slates and clocks were not provided to any reduced-cost schools in 2013, but randomly provided to half of them in 2014. Kerwin and Thornton (2021) discuss and quantify the differences between the full- and reduced-cost program versions.

[10] In schools with fewer than 100 students, all available P1 students were included in the study. In the original 38 schools, 40 P1 students were sampled at the beginning of the school year with an additional 60 sampled at the end of the year; in the 90 additional schools 80 P1 students were sampled at the beginning of the year and 20 at the end of the year. Students sampled at the end of the year are about half a year older and 1.70 percentage points more likely to be female. There are no differences in attendance at the end of the school year by treatment arm; the average treatment effect of the program also does not vary substantially by when a student was sampled (Buhl-Wiggers et al., 2018a).

moderators (measured in P1, at the beginning of 2014), except for baseline test scores. Because top-up sample students lack baseline test scores, we recode missing values to zero and include an indicator for missing values.[11] Our main analysis includes 4,868 students (1,427 in the control group, 1,681 in the full-cost treatment group and 1,760 in the reduced-cost treatment group). See Appendix Table A1 for details on the construction of the main analysis sample.

3.3. Learning outcomes

Our outcome measure captures reading performance in Leblango at the end of P3 in 2016, specifically an index of scores on the Early Grade Reading Assessment (EGRA).[12] The EGRA is an internationally standardized exam—externally validated in Leblango (RTI International, 2019). 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 Kerwin and Thornton (2021), we construct a principal component score index 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. Treatment effects on test scores at the end of P3 reflect exposure to NULP for three years, in P1, P2, and P3.[13]

3.4. Covariates

We have student-, teacher-, and school-level variables. Student-level covariates comprise baseline test scores, an indicator for a missing baseline score, a male indicator, and student age at baseline (censoring the data at 7 and 13, affecting just 0.39 percent of all observations).[14] Teacher-level covariates include a male indicator, age, years of teaching experience, and years of completed schooling.[15] School-level variables include four variables that were collected in the year prior to the schools entering the study: total P1–P7 student enrollment, the total number of P1–P7 teachers, the pupil-to-teacher ratio in grades P1–P3, and the per-capita number of passes on the Primary Leaving Exam (PLE) in the year before the intervention started (2012 for the first 38 schools that joined the sample in 2013, and 2013 for the remaining 90 schools that entered the sample in 2014).[16] We also include 14 variables that are based on data from 2017, such as the share of students who attended nursery school and an index of access to books.[17]

Appendix Table A2 presents covariate means by study arm along with balance tests. By construction, 50 percent of the students in the sample are girls; the average student is between eight and nine years old. Teachers average 39 to 43 years of age with 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 about 0.05 PLE passes per capita. 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.

4. Average treatment effects

We begin by estimating the average effect of each version of the NULP in our main analysis sample with the following linear model:

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

where Y_{isc} denotes the Leblango reading index for student i in school s in stratification cell c. FC_s and RC_s indicate assignment to the full-cost and reduced-cost treatment arms for school s, respectively. β_{FC} and β_{RC} represent the average treatment effects of the two program versions. $Y_{i,P1}$ denotes the baseline P1 Leblango reading test score index and D_i is an indicator equal to one when the baseline score is missing. α_c is a stratification cell fixed effect and ϵ_{isc} is a mean-zero term that captures the effects of all omitted determinants of test scores. We cluster standard errors at the school level given the school-level treatment assignment.

[11] Because students have no background in reading when they arrive at school, almost all students who took the test (87 percent), received a zero on their 2014 baseline test score.

[12] Only students present in school were administered the endline exam; See Appendix Table A1.

[13] More precisely, we test students in their third year, which means P3 for 80 percent of the students we test. The remainder repeated an earlier grade.

[14] All student-level moderators are measured at the beginning of 2014.

[15] Each student was exposed to up to three different teachers over the three years of evaluation. We use only the characteristics of students' teachers in 2016, choosing a single set of teachers to reduce the dimensionality of the data; these are the most-recent teachers, and so have the most-proximal influences on their 2016 endline test scores. They also have the fewest missing values.

[16] PLE passes per capita are defined as the total number of students who passed the PLE at the end of P7 divided by total enrollment from P1 to P7. We do this because there are high dropout rates between P6 and P7 as schools strategically try to raise the fraction of P7 students who pass the exam (Gilligan et al., 2019), and because we did not collect enrollment for P7 separately.

[17] Because these data were collected post-treatment, we include only variables unlikely to have been affected by the treatment; the balance tests presented in Appendix Table A2 reveal no meaningful differences across study arms for these variables.

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
Sampling Strata FE		Yes	Yes	Yes

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

Table 1 presents the estimates from four versions of Eq. (1). Column (1) shows unconditional treatment effects (i.e. the simple mean difference). Column (2) adds stratification cell fixed effects, α_c, for consistency, because a few stratification cells have different shares of schools in each study arm; including these fixed effects also improves the statistical efficiency of our estimates (Bruhn and McKenzie, 2009). Column (3) adds controls for students' baseline test scores, dropping students with a missing test score. Column (4) keeps students with missing baseline scores and includes an indicator for missing values.[18]

Using our preferred specification in Column (4), our estimated average treatment effects are 1.40 SDs for the full-cost treatment and 0.74 SDs for the reduced-cost treatment. 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 Column (4) as our main specification for the remainder of the paper.

These estimates represent very large impacts. The full-cost program effect sits at the 99th percentile of the overall distribution of impacts of the primary-school education programs in McEwan (2015) across all outcome measures; not a single program in McEwan has such a large effect on reading scores. Even the reduced-cost program effects are large relative to the literature; 95 percent of the experiments in McEwan yield treatment effects below 0.45 SDs, with the average being 0.10 SDs.

5. Establishing treatment effect heterogeneity

This section presents evidence of treatment effect heterogeneity using the classical statistical bounds that rely only on the information in the marginal outcome distributions.

5.1. Formalities and implementation

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 specific marginal distributions, without the addition of any further identifying information. In the context of our three-armed experiment, treatment represents either the full-cost or reduced-cost version of the NULP.

For continuous variables, the Fréchet–Höffding bounds are:

$$max[F_1(Y_1|D=1) + F_0(Y_0|D=1) - 1, 0] \leq F(Y_1, Y_0|D=1) \leq min[F_1(Y_1|D=1), F_0(Y_0|D=1)] \quad (2)$$

where $F_1(.)$ is the marginal distribution of the outcome variable under treatment and $F_0(.)$ is the marginal distribution under control. The lower bound corresponds to the case of perfect negative dependence or "rank inversion" as it implies a

[18] 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.

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.[19]

To estimate these bounds, we collapse the outcome distributions for the treatment and control arms into percentiles—to simplify the computations and because the three arms contain different numbers of students. Subtracting the control outcome from the treated outcome for a given percentile yields the treatment effect for that percentile at the FH upper bound. A similar operation with the control outcome percentiles inverted provides the treatment effects associated with the FH lower bound distribution.[20]

Cambanis et al. (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, $var(Y_1 - Y_0)$, and thus the impact standard deviation.[21] [22]

We also present the quantile treatment effects associated with the 5th, 25th, 50th, 75th, and 95th percentiles of the control-group outcome distribution. For each combination of upper or lower bound and full- or reduced-cost program, we calculate several additional statistics. First, we calculate the standard deviation of the estimated treatment effects (the impact standard deviation) as the square root of the variance of the percentile-specific impacts.[23] Second, we calculate the Pearson correlation between the percentiles of the treated and control outcome distributions. Third, we estimate the fraction of students with a positive impact as the fraction of non-negative percentile-specific impact estimates. The fraction positive is not super-additive and thus need not fall into the range defined by our bounds. We compute bias-corrected confidence intervals using the non-parametric bootstrap, drawing 1,000 bootstrap samples of students from the main analysis sample, clustered by school and stratified by stratification cell.[24]

5.2. Findings

Table 2 presents estimates of the statistics associated with the FH bounding distributions. Columns (1) and (3) give statistics under rank preservation (the FH upper bound distribution); Columns (2) and (4) present statistics under rank inversion (the FH lower bound distribution). We focus on the bottom three rows of the table and defer discussion of treatment effects at particular quantiles of the outcome distribution (which under rank preservation correspond to the QTEs) to Section 6.2.

We begin with the impact standard deviation, finding (huge!) bounds of (1.07, 2.62) for the full-cost program and of (0.64, 2.22) for the reduced-cost program. The lower bound for the full-cost program equals 76 percent of its mean impact in Column (4) of Table 1, while the lower bound for the reduced-cost program equals about 87 percent of its mean impact. If treatment effects are normally distributed then 29 percent of students in the full-cost program have impacts of at least 2 SDs and more than 10 percent have negative impacts. Another way of looking at the bounds of the treatment effects compares the variation *within* the full-cost program to the variation in average treatment effects *across* the full- and reduced-cost programs. Again assuming normality, the difference between the 5th and 95th percentile full-cost treatment effects equals 3.50 SDs—over four times the difference in average impacts between the full- and reduced-cost programs. The difference in effects within the full-cost program also far exceeds the difference in average treatment effects between the most- and least-effective programs among the 76 randomized experiments covered in McEwan (2015), which vary in mean impact from −0.57 to 1.51 SDs, a range of 2.08 SDs.

These bounds apply to our outcome based on the EGRA. We also care about bounds on the distribution of NULP impacts on reading ability, which the EGRA measures only with error. The literature offers surprisingly little evidence on the extent of measurement error in the EGRA. For the Spanish-language EGRA, the test–retest reliability varies from 0.6 to 0.8 across the test modules we use (Jiménez et al., 2014). To get a sense of the potential bounds on impacts on reading ability, in Appendix Figure A1 we present the results of a simulation exercise in which we add classical (i.e., mean-zero, normally

[19] 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 90th percentile of the full-cost program outcome distribution equals the 90th percentile of the control outcome distribution. In contrast, with rank inversion the counterfactual for a student at the 90th percentile of the full-cost program outcome distribution equals the 10th percentile of the control group outcome distribution.

[20] 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.

[21] To see the intuition, suppose that $F_1 \sim U[0, 1]$ and $F_0 \sim U[0, 1]$, i.e. both 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, the FH lower bound distribution, with its attendant rank inversion, implies treatment effects that 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).

[22] While HSC (1997) are correct when they state that "[t]hese 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).

[23] Calculating the impact standard deviation using the percentiles rather than some finer approximation to the outcome distributions likely leads to a mild understatement of the true population bounds.

[24] We do not sample schools and then students within schools in our bootstrap for computational simplicity. As a result, we may understate the sampling variability in our estimates. The intra-class correlation coefficient of our outcome is 0.23, suggesting only a modest effect on our estimates.

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				
5th	0.034	5.631	0.034	4.553
	[0.034,0.057]	[5.407,5.831]	[0.044,0.046]	[4.314,4.802]
25th	0.386	3.318	0.205	2.199
	[0.285,0.489]	[2.961,3.590]	[0.171,0.266]	[1.981,2.503]
50th	1.333	1.333	0.619	0.619
	[1.014,1.701]	[1.014,1.701]	[0.462,0.828]	[0.462,0.828]
75th	2.577	−0.355	1.458	−0.536
	[2.149,2.891]	[−0.596,−0.171]	[1.125,1.807]	[−0.756,−0.371]
95th	2.964	−2.633	1.886	−2.633
	[2.518,3.338]	[−3.049,−2.383]	[1.503,2.262]	[−3.049,−2.382]
Impact Standard Deviation	1.066	2.615	0.642	2.218
	[1.019,1.107]	[2.586,2.645]	[0.609,0.686]	[2.195,2.240]
Outcome Correlation	0.932	−0.655	0.975	−0.577
	[0.907,0.959]	[−0.698,−0.612]	[0.955,0.989]	[−0.610,−0.541]
Fraction Positive	0.980	0.697	0.980	0.646
	[0.960,0.980]	[0.657,0.707]	[0.960,0.980]	[0.606,0.667]

Notes: Columns (1) and (3) show statistics estimated using the Fréchet–Höffding lower-bound distribution from Eq. (2), while Columns (2) and (4) use the upper-bound distribution. We construct the bounding distributions as described in Section 5.1. All estimates use the main analysis sample. Bias-corrected confidence intervals, bootstrapped using 1,000 replications, in brackets.

distributed) measurement error to our outcomes for one or both treatment arms. As expected, adding measurement error to only the treatment group outcomes modestly increases the FH lower bound on the impact variance, while adding it only to the control group outcomes modestly reduces it. Adding it to both yields a very small decrease. Our prior is that there is measurement error in both arms, with perhaps a bit less in the treatment arm. The simulations imply that in this case, even with implausibly large amounts of measurement error our qualitative conclusions carry over to impacts on underlying reading ability.

For both the full- and reduced-cost treatments, rank preservation (by construction) yields large positive Pearson outcome correlations, while rank inversion yields large negative ones. Our bounds on the fraction with a positive treatment effect illustrate the underlying intuition of the FH bounds. Consider the example in which both Y_1 and Y_0 have $U[0, 1]$ distributions. In this case, rank preservation yields a fraction positive (more precisely, non-negative) of 1.00 because $Y_1 = Y_0$, so all of the treatment effects equal zero. In contrast, rank inversion yields a fraction positive of 0.50, as the bottom half of the treated units get linked to the top half of the untreated units and vice versa. More generally, as long as the two distributions share some common support, rank inversion necessarily leads to at least some fraction of the treated units having negative treatment effects.[25]

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 the fraction only falls to about 0.70 for the full-cost program and 0.65 for the reduced-cost program.

5.3. Testing the null of a common treatment effect

In the preceding section, we carefully avoided performing a simple hypothesis test of the null of a zero impact standard deviation based on the bootstrapped confidence intervals for the estimated lower-bound impact standard deviations. 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, standard deviations must, by construction, lie in the interval $[0, \infty)$. Our null of zero lies on the edge of that set.[26] HSC (1997, Appendix E) makes a strong case that the bootstrapped confidence intervals, 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. Appendix Table A3 repeats a subset of their analysis using our data with the same qualitative conclusion.

[25] To see this, first change the example so that the treated unit outcomes are distributed $U[0.90, 1.90]$. 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 $U[1.10, 2.10]$ implies a fraction positive of 1.00 even under rank inversion, because every treated outcome with positive support exceeds every control outcome with positive support.

[26] To see the problem at a very prosaic level, think about a sample from an RCT where the null holds for some outcome. Imagine calculating the impact standard deviation using the sample (as we do above). 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 this issue by using the randomization inference procedure developed in Appendix E of HSC (1997).[27] 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.[28] The results are presented in Appendix Table A4. Comparing the impact standard deviation lower bounds from Table 2 with the cutoff values of Appendix Table A4, we can easily reject the null with a p-value of 0.0001 for both program versions.[29]

6. Exploring treatment effect heterogeneity

This section examines the extent to which additional assumptions – stochastic increasingness and rank preservation – reduce and clarify the variation revealed by the classical bounds.

6.1. FL bounds

6.1.1. Introduction

The FH bounds on the standard deviation of $(Y_1 – Y_0)$ imply a great deal of treatment effect heterogeneity. In this section, we consider alternative bounds developed in FL (2021). They show that limiting consideration to joint distributions of potential outcomes that exhibit the property of "mutual stochastic increasingness" (MSI) allows for informative pointwise bounds on the average treatment effects at specific quantiles of the potential outcomes, as well as on the fraction of students who experience negative treatment effects.

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" (FL 2021). This means that if student A has a higher test score than student B under the status quo (e.g., in the control), student A will also probably have a higher score than student B if exposed to the treatment, and similarly if their roles were reversed. MSI implies a positive rank correlation,[30] but a positive rank correlation does not imply MSI.[31] 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, a far weaker restriction.[32]

Does MSI make sense in our substantive context? 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. Latent ability and effort likely imply better performance in both the treated and untreated states in our context. We can think of them in terms of a one factor ("ability") model with noise, a model we find quite plausible in our context. At the same time, we worry that our data contains students who would flourish in the control world of call-and-response in English and flail in the NULP world of scripts and slates in Leblango, or the reverse. Too many such students imply that MSI fails even as an approximation.

6.1.2. Formalities and implementation

FL (2021) define the potential outcomes Y_1 and Y_0 as mutually stochastically increasing if the following property holds:

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

This means, 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, FL (2021) show that the lower-bound CDF is given by

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

[27] 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.

[28] Buhl-Wiggers et al. (2020) provide the details of our implementation of the test.

[29] We also test the null hypothesis of a common treatment effect using the tests suggested by Chernozhukov and Fernández-Val (2005) and Chung and Olivares (2021). Appendix Tables A5 and A6 show that we reject the null with both tests.

[30] This property links our analysis to Tables 5A and 5B in HSC (1997), which describe the distributions of impacts randomly sampled conditional on particular values of the rank correlation between Y_1 and Y_0.

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

[32] To see that MSI is less restrictive than rank preservation from another angle, suppose that $Y_1 = Y_0 + v$, where Y_0 is continuous and v has a symmetric, continuous distribution independent of Y_0. This setup satisfies MSI but does not satisfy rank preservation due to the random component.

where $\Delta = Y_1 - Y_0$. The upper bound is given by

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

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 (2021) algorithm for this proceeds as follows: First, compute $F_0(y+t)$ as the sample mean of the indicator $\mathbb{1}(Y_i \leq y+t)$ in the control-group data. Similarly, compute $F_1(y+t)$ as the sample mean of the indicator $\mathbb{1}(Y_i \leq y+t)$ in the treatment-group data. Then plug those estimates into Eqs. (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:

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

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

where $d \in \{0, 1\}$.

Intuitively (though not obviously), the pointwise lower bound on the conditional treatment effect in Eq. (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 Eq. (6) has rank preservation below the evaluation point and independence above it.

6.1.3. Findings

We present the pointwise FL bounds on the conditional expected impacts in Table 3. Panel A presents the lower and upper bounds for the average treatment effects of each program by control-group percentile. The mean effects of the full-cost program range from 0.20 SDs to 2.65 SDs for the 5th percentile student, and from −0.57 SDs to 4.25 SDs for the 95th percentile student. For the reduced-cost program, the 5th percentile student on average gains between 0.16 and 1.92 SDs, and the 95th percentile student sees mean effects ranging from a 1.19 SD *loss* to a 3.23 SD gain. The upper bounds increase nearly monotonically with percentiles of the control-group outcome distribution for both programs, while the lower bounds initially rise and then fall for the highest percentiles. Panel B shows bounds on the fraction of students with negative treatment effects at each control-group percentile. The lower bound on the fraction negative is always zero, while the upper bound increases monotonically with test score percentiles under control status, reaching 0.70 for the 95th percentile for the full-cost program and 0.84 for the reduced-cost program.

Imposing MSI provides some valuable substantive insight. First, all of the bounds are quite wide: the average full-cost program treatment effect for a student at the median of the control-group 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, 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.

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 comes through quantile treatment effects (QTEs).[33] 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 difference in quantiles between the treatment- and control-group outcome distributions.

QTEs admit 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 the treatment increases the outcome variance. Graphing the QTEs against the percentiles can add meaningfully to the information provided by the average treatment effect.[34]

[33] Koenker and Bassett (1978) began the literature on quantile regression in economics. Important early applications in program evaluation include Lehmann and D'Abrera (1975), and Doksum (1974) in the statistics literature and HSC (1997), Koenker and Bilias (2001), Abadie et al. (2002), and Bitler et al. (2006) in economics. HSC (1997) do not use the term QTE because it had not yet entered the applied econometric lexicon when they wrote.

[34] Indeed, it surprises us that such graphs have not become routine in experimental evaluations!

Table 3
Frandsen and Lefgren bounds on treatment effects by percentile.

	Full-cost Program		Reduced-cost Program	
Percentiles under control status	Lower Bound (1)	Upper Bound (2)	Lower Bound (3)	Upper Bound (4)
A) Bounds on Average Treatment Effect				
5th	0.195	2.652	0.163	1.919
	(0.057)	(0.057)	(0.065)	(0.047)
25th	0.307	3.015	0.183	2.160
	(0.029)	(0.071)	(0.039)	(0.058)
50th	0.565	3.774	0.263	2.732
	(0.049)	(0.092)	(0.030)	(0.083)
75th	0.600	4.269	0.085	3.140
	(0.052)	(0.104)	(0.032)	(0.100)
95th	−0.573	4.246	−1.193	3.233
	(0.050)	(0.164)	(0.039)	(0.185)
B) Bounds on Fraction with Negative Treatment Effect				
5th	0.000	0.289	0.000	0.296
	(0.000)	(0.042)	(0.000)	(0.043)
25th	0.000	0.310	0.000	0.341
	(0.000)	(0.027)	(0.000)	(0.027)
50th	0.000	0.347	0.000	0.459
	(0.000)	(0.020)	(0.000)	(0.021)
75th	0.000	0.465	0.000	0.648
	(0.000)	(0.016)	(0.000)	(0.017)
95th	0.000	0.696	0.000	0.843
	(0.000)	(0.012)	(0.000)	(0.011)

Notes: In Panel A, Columns (1) and (3) present estimates of Eq. (5) for the full- and reduced-cost programs respectively; Columns (2) and (4) present estimates of Eq. (6). In Panel B, Columns (1) and (3) use Eq. (3) to estimate the lower bound on the fraction of students with negative treatment effects, and Columns (2) and (4) use Eq. (4) to estimate the upper bound on the fraction of negative treatment effects. All estimates use the main analysis sample. Bootstrapped standard errors, clustered by school and computed using 100 replications, in parentheses.

The second interpretation presumes rank preservation and returns us to the world of the FH upper-bound distribution. In this interpretation, the QTEs represent impacts *at quantiles* as well as *on quantiles*.[35] Thus, we can make statements such as "the treatment improves the test score of the X^{th} percentile student by Y SDs". The first interpretation does not allow such statements, because without rank preservation, the joint distribution could be anything (consistent with the given marginals).[36]

6.2.2. Implementation

We estimate QTEs for 19 ventiles (i.e. every fifth percentile from the 5th to the 95th) using the estimator defined in Koenker and Bassett (1978), embodied in Stata's `qreg` command.[37] We present bootstrapped standard errors based on 250 replications, resampling schools from within their original stratification cells. 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; these average effects correspond most closely to our QTEs, which also do not include any control variables.[38]

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: the equality of the QTEs at the same ventiles for which we present estimates. This equality 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. Rejections using this test statistically imply a non-zero impact variance but failure to reject does not imply an impact variance of zero.

6.2.3. Findings

Fig. 1 shows the quantile treatment effect estimates. Both NULP variants exhibit monotonically increasing treatment effects across the quantiles of the outcome distribution. We see no effect of the two programs on the 5th percentile of

[35] For example, under rank preservation, the 75th percentile QTE gives the impact on students at the 75th percentile of the untreated outcome distribution.

[36] Bitler et al. (2014) compare the knowledge produced by quantile treatment effects and by subgroup impacts defined based on baseline outcomes.

[37] 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).

[38] Including conditioning variables when estimating a QTE changes the substantive meaning of the estimand, which becomes a conditional QTE. Powell (2020) shows how to compute unconditional QTEs while including covariates under an assumption of conditional rank similarity. While we doubt this condition holds in our setting (see Section 6.2.4 below), Appendix Figure A2 shows that applying his method to our data does little to change the findings.

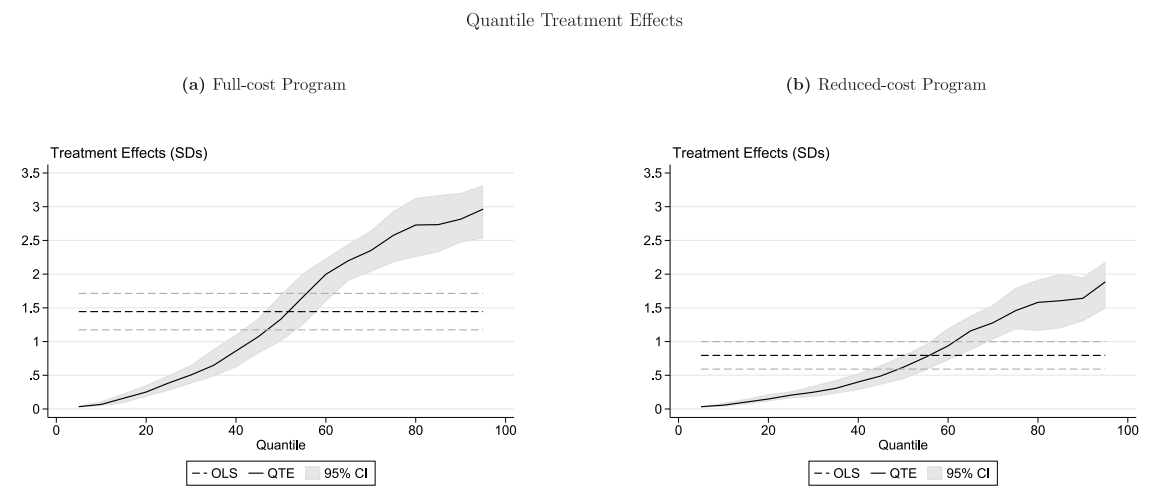

Fig. 1. Quantile Treatment Effects. *Notes*: Quantile treatment effects are estimated via the Koenker and Bassett (1978) method as described in Section 6.2.2, using the main analysis sample. Outcome is the Leblango reading test score index standardized with respect to the control group. Solid lines are quantile treatment effect estimates; gray regions are bootstrapped 95% confidence intervals. The dark dashed line is the average treatment effect estimated via Eq. (1), with the 95% confidence interval indicated via light dashed lines.

outcomes.[39] The QTEs increase steadily up to about 2.97 SDs on the 95th 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 the NULP strongly increase the variance of academic outcomes as well as the mean. Rank preservation adds the further implication that students who would struggle under the existing regime would also struggle with the NULP, while students who do well in the control state would also do well under the NULP. As shown in Appendix Table A7, we reject the null of equal QTEs at the 0.001 level for both program variants.

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 et al. (BGH) (2005) point out that rank preservation does have testable implications.[40] Under rank preservation, characteristics of units not affected by treatment should look the same at corresponding quantiles of the treatment and control outcome distributions. As with our test of equal QTEs, because we test an implication of the null 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 BGH (2005). First, we divide our outcome (i.e. Leblango reading score index) into quartiles separately by treatment arm.[41] Within each quartile, we regress each of our student-, teacher-, and school-level covariates on indicators for each of the two treatments and stratification cell fixed effects.[42] The coefficients on the treatment indicators represent the quartile-specific mean differences in the covariate under rank preservation, and should equal zero up to sampling variation.[43]

Table 4 presents the results, where we easily reject rank preservation. With 48 tests (12 covariates and 4 quartiles), at the 10 percent level we would expect a total of about five rejections for independent tests; our tests are not independent

[39] Nearly 10 percent of the control group scores zero on the entire Leblango EGRA while the 5th percentile scores in the two treatment arms differ only marginally from zero. In one sense, this is a "floor" effect; in another sense, it clearly indicates that these students have learned very little after three years as they are unable to recognize even a single letter of the alphabet.

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

[41] The choice of quartiles, rather than, say, quintiles, 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 because a test at a specific quantile would have no power. Increasing the width of the test interval increases statistical power while weakening the correspondence between the null implicit in the test and the null of covariate balance at specific quantiles.

[42] We modify the procedure in BGH by adding a step in which we subtract off the overall average effect of each treatment (across all four quartiles) on the particular covariate. This focuses the test on changes in ranks by removing the small amounts of overall imbalance that result from sampling variation.

[43] See our working paper, Buhl-Wiggers et al. (2020), for details on the construction of the standard errors.

Table 4
Tests of covariate balance by endline test score quartile.

	Full-cost program				Reduced-cost program			
	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
	[-0.031;0.028]	[-0.033;0.031]	[-0.041;0.041]	[-0.074;0.075]	[-0.031;0.029]	[-0.033;0.035]	[-0.042;0.042]	[-0.074;0.072]
1(BL Missing)	0.002	−0.065*	0.043	0.040	0.024	−0.079**	0.016	0.037
	[-0.054;0.055]	[-0.055;0.062]	[-0.056;0.059]	[-0.057;0.054]	[-0.054;0.060]	[-0.057;0.062]	[-0.057;0.059]	[-0.055;0.054]
1(Student Male)	0.028	−0.054	−0.027	0.043	0.077**	−0.018	−0.076**	0.053
	[-0.061;0.060]	[-0.061;0.059]	[-0.060;0.062]	[-0.057;0.059]	[-0.058;0.059]	[-0.060;0.056]	[-0.058;0.053]	[-0.057;0.060]
Student's Age	−0.025	−0.114	0.179**	−0.046	0.002	−0.131	0.107	−0.018
	[-0.145;0.132]	[-0.151;0.141]	[-0.129;0.124]	[-0.123;0.126]	[-0.134;0.137]	[-0.137;0.148]	[-0.126;0.122]	[-0.121;0.123]
1(Male Teacher)	−0.091***	−0.024	0.012	0.048**	−0.002	−0.004	−0.035	0.005
	[-0.039;0.039]	[-0.041;0.040]	[-0.044;0.043]	[-0.036;0.037]	[-0.037;0.038]	[-0.040;0.040]	[-0.043;0.042]	[-0.036;0.036]
Teacher's Age	−1.077***	0.216	−0.048	0.549	−2.470***	0.024	0.729*	0.289
	[-0.568;0.648]	[-0.610;0.631]	[-0.650;0.637]	[-0.588;0.581]	[-0.615;0.631]	[-0.653;0.601]	[-0.649;0.652]	[-0.565;0.575]
Teacher's Experience	0.063	−0.090	−0.269	−0.310	−0.974***	−0.120	0.507	−0.334
	[-0.485;0.534]	[-0.525;0.524]	[-0.540;0.560]	[-0.528;0.562]	[-0.480;0.522]	[-0.571;0.496]	[-0.555;0.561]	[-0.499;0.536]
Years of Education	−0.145***	0.027	0.013	0.116**	−0.081	0.013	0.098*	−0.031
	[-0.087;0.088]	[-0.104;0.104]	[-0.097;0.105]	[-0.097;0.092]	[-0.081;0.086]	[-0.099;0.103]	[-0.092;0.097]	[-0.088;0.093]
School's Enrollment	−5.341	−9.405	−4.079	−5.977	17.772**	4.160	−1.445	−16.225*
	[-14.100;16.739]	[-17.202;16.389]	[-17.126;18.988]	[-15.166;14.331]	[-14.612;14.688]	[-17.750;17.623]	[-18.367;17.639]	[-15.497;13.936]
Pupil–Teacher-Ratio	1.769***	0.767	−3.431***	−1.172*	−0.511	−0.178	−0.806	0.741
	[-1.024;1.193]	[-1.234;1.199]	[-1.363;1.304]	[-1.025;1.010]	[-1.038;1.069]	[-1.292;1.200]	[-1.247;1.336]	[-1.046;0.975]
PLE Passes per Capita	0.334***	0.121	0.158	−0.291***	0.712***	0.072	−0.010	−0.341***
	[-0.169;0.139]	[-0.151;0.143]	[-0.166;0.161]	[-0.112;0.122]	[-0.161;0.138]	[-0.161;0.153]	[-0.165;0.158]	[-0.113;0.112]
Number of Teachers	−0.157	−0.120	0.583***	0.163	0.479***	0.212*	0.080	−0.259**
	[-0.183;0.181]	[-0.189;0.188]	[-0.193;0.217]	[-0.176;0.162]	[-0.186;0.184]	[-0.202;0.192]	[-0.214;0.212]	[-0.163;0.156]

Notes: Bitler et al. (2005)/Djebbari and Smith (2008) rank preservation tests, implemented as described in Section 6.2.4 and using our 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 before taking the within-quartile differences. Each column presents differences for the indicated quartile. Bootstrapped 90% confidence intervals in brackets: * $p < 0.01$; ** $p < 0.05$; *** $p < 0.1$.

which implies that we should expect even fewer rejections.[44] We reject the null at the 10 percent level in 29 percent of tests for the full-cost program, and in 27 percent of tests for the reduced-cost version.[45]

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

Using only students with non-missing values of baseline test scores, Fig. 2 plots the test score transitions within treatment arm by quartile; the high fraction of students with zero baseline scores forces us to combine the bottom two quartiles. The figure shows 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. While students who start out in the top quartile have a higher probability of ending up in the top quartile within their study arm in all three arms, their advantage is quite modest. The same finding holds for the third quartile (not shown). Overall, the evidence in Fig. 2 indicates either a very noisy test, the failure of the one-factor model, or both. We have a high degree of faith in the EGRA as a measure of basic reading ability, and so a very noisy test seems unlikely.

The covariate balance tests in Table 4 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 Fig. 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 status quo (control-group) test score distribution.

7. Systematic treatment effect variation

Having provided strong evidence of meaningful essential heterogeneity and examined whether and what we can learn about that heterogeneity under the substantive assumptions of mutual stochastic increasingness and rank preservation, we now investigate the extent to which the treatment effect heterogeneity we observe correlates with observed covariates.

44 Appendix Table A8 shows qualitatively similar results without the stratification cell fixed effects.

45 We also conduct tests of the null of rank similarity, the stochastic analogue of rank preservation, due to Dong and Shen (2018) and Frandsen and Lefgren (2017). Their tests build on the same broad intuition as BGH (2005) that covariates (and distributions of covariates) should balance between treated and untreated units at the same quantiles of their respective outcome distributions under rank preservation or rank similarity. Appendix Figure A3 presents our results from implementing the Dong and Shen (2018) tests as in Cummins (2017) and Appendix Table A9 presents our Frandsen and Lefgren (2017) results. Both alternative testing strategies reject the null at least as strongly as our BGH (2005) tests.

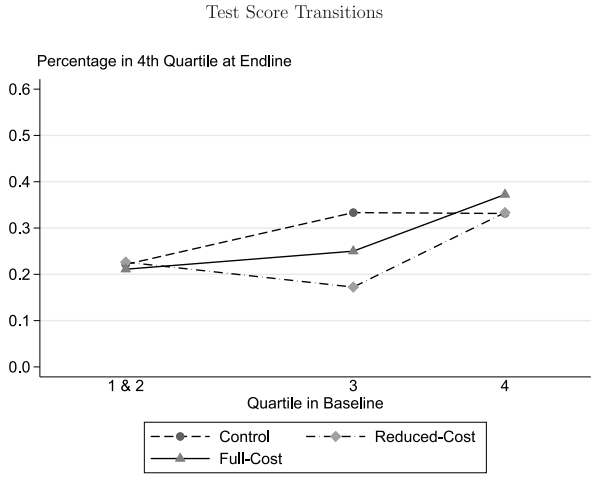

Fig. 2. Test Score Transitions. *Notes*: Sample is 2,395 students with a complete set of moderators and baseline results (723 control, 849 full-cost treatment, 823 reduced-cost treatment). Graph presents the share of students from each quartile of baseline scores who end up in the fourth quartile of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

We follow (some of) the literature in labeling these variables "moderators".[46] Following Djebbari and Smith (2008), we divide the extant treatment effect heterogeneity into a "systematic" component – the part that the moderators capture – and an "idiosyncratic" component, while keeping in mind that this division depends on the set of available candidate moderators. In our (philosophical) view, if we measured all possible moderators we could convert all of the essential heterogeneity into systematic heterogeneity.

Measuring and identifying systematic treatment effect heterogeneity can help identify "what works for whom", which aids in targeting interventions toward those most likely to benefit from them.[47] Given the strong average treatment effects of both versions of the NULP program, learning about predictors of relatively low (or even negative) treatment effects could allow compensatory action within classrooms and motivate further study of particular aspects of program implementation with an eye towards improving the treatment effects of those who benefit the least. Teachers whose characteristics predict lower average treatment effects for their students could receive further support 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 for broader reasons, as with girls or ethnic minorities in primary and secondary school in much of the developing world (and indeed in the US as well).

7.1. Candidate moderators

The design of the NULP evaluation did not have identification of effect moderators 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 the effects of modest but substantively meaningful moderators.[48] As above, we group our candidate moderators into three sets: student characteristics, teacher characteristics, and school characteristics.

Theory and 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) suggest that the productivity of additional inputs depends on previous investments ("dynamic complementarities"), and extensive research has shown that students' initial ability matters a great deal for learning trajectories (Banerjee et al., 2017). This motivates our inclusion of baseline test scores. Program impacts may vary with student age in a context like northern Uganda wherein students start school at different ages and many students repeat grades. 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, student-to-teacher ratios predict student learning and may affect how well teachers can implement interventions. School size may capture economies of scale.

Policy interest drives the inclusion of some other candidate moderators. There is considerable demand from policymakers for evidence on interventions that work well for girls—see, e.g. Evans and Yuan (2019) as well as evidence in

[46] 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 adult outcomes. Interpreted that way, we investigate moderators for our mediator.

[47] See, e.g. Berger et al. (2001), Bitler et al. (2017), Lee and Shaikh (2014), and Weiss et al. (2014).

[48] See, e.g. Gelman (2018) on sample size requirements in moderation analyses.

Lim and Meer (2017) that assigning girls to female teachers improves their test scores. These factors help motivate the inclusion of student and teacher sex. We also include PLE passes per capita[49] as a candidate moderator, since both administrators and parents commonly use the pass rate as a proxy for the quality of Ugandan primary schools and it is routinely collected and readily available.

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 moderators (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.[50] 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 and teachers 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.[51] In our context, this yields the following linear model:

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

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- or reduced-cost treatment arm, respectively, with associated coefficients β_{FC} and β_{RC}. We let X_i^j denote the value of moderators $j \in 1, \ldots, J$ for student i. Because we have just 128 schools in our analysis, we conduct our conventional analyses without including any of the school-level moderators and only use student- and teacher-level variables.[52]

We de-mean all the moderators prior to inclusion so that β_{FC} and β_{RC} retain their interpretation as average treatment effects. The coefficients β_{FC}^j and β_{RC}^j indicate the conditional expected change in the relevant treatment effect for a one-unit change in moderator j, 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 students to 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_{FC}^j and β_{RC}^j. For example, if X_i^1 indicates that a student is female, 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 et al. (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

Table 5 presents 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), teacher characteristics in Column (3), and (our preferred specification) both sets of characteristics in Column (4). We find only limited evidence of systematic variation in treatment effects. Students with missing baseline scores have smaller treatment effects, but this effect attains only marginal statistical significance (in a table full of hypothesis tests) and presumably represents not a causal moderation effect but instead the missing test score acting as a proxy for some other student characteristics not

[49] As above, this is the ratio of passes to the total number of students in the school.

[50] We assess the underlying dimensionality of our set of candidate moderators via a principal components analysis. Appendix Table A10 reveals that the moderators do not, for the most part, measure overlapping constructs: the most important principal component explains just 14 percent of the overall variance.

[51] Depending on the available sample size and the size and nature of the set of candidate moderators, the set of included moderators in a particular study may include all available candidate moderators, or some subset chosen in an ad hoc manner to avoid multi-collinearity and/or over-fitting and/or concerns about multiple hypothesis testing.

[52] When we run a school-level regression of test scores on all the school-level moderators and their interactions with the treatment indicators, the R^2 is 0.93. Adding the school-level averages of moderators measured at the student and teacher levels raises the R^2 to 1.

Table 5
Systematic variation in treatment effects.

	Base Model	Covariates of:		
		Students	Teachers	All
	(1)	(2)	(3)	(4)
Full-cost program	1.396***	1.337***	1.325***	1.274***
	(0.116)	(0.143)	(0.131)	(0.156)
Reduced-cost program	0.738***	0.719***	0.791***	0.792***
	(0.108)	(0.133)	(0.102)	(0.128)
Full-cost*Baseline Test Score		0.085		0.097
		(0.135)		(0.138)
Reduced-cost*Baseline Test Score		0.069		0.066
		(0.145)		(0.146)
Full-cost* 1(BL Missing)		−0.248*		−0.235*
		(0.130)		(0.126)
Reduced-cost* 1(BL Missing)		−0.148		−0.153
		(0.105)		(0.101)
Full-cost* 1(Male)		−0.037		−0.056v
		(0.103)		(0.106)
Reduced-cost* 1(Male)		−0.102		−0.106
		(0.097)		(0.097)
Full-cost*Age		0.076		0.070
		(0.059)		(0.057)
Reduced-cost*Age		0.032		0.016
		(0.058)		(0.055)
Full-cost* 1(Male Teacher)			0.379	0.362
			(0.255)	(0.248)
Reduced-cost* 1(Male Teacher)			−0.282	−0.300
			(0.247)	(0.244)
Full-cost*Teacher's Age			−0.014	−0.014
			(0.033)	(0.033)
Reduced-cost∗Teacher's Age			0.000	−0.002
			(0.031)	(0.031)
Full-cost*Teacher's experience			−0.000	0.000
			(0.035)	(0.035)
Reduced-cost*Teacher's experience			0.005	0.005
			(0.030)	(0.030)
Full-cost*Years of Education			−0.038	−0.036
			(0.104)	(0.104)
Reduced-cost*Years of Education			−0.028	−0.025
			(0.090)	(0.088)
Observations	4,868	4,868	4,868	4,868
R-squared	0.179	0.181	0.186	0.188
Adj-R-Squared	0.170	0.171	0.175	0.176
Group*Year*Cohort FE	Yes	Yes	Yes	Yes

Notes: Estimates of Eq. (7) using the main analysis sample. Outcome is the Leblango reading score, standardized with respect to the control group. Each specification also includes main effects for all of the covariates that are interacted with the treatment indicators. All interacted covariates are de-meaned prior to constructing the interaction terms. Heteroskedasticity-robust standard errors, clustered by school, in parentheses: * $p < 0.01$; ** $p < 0.05$; *** $p < 0.1$.

included among our candidate moderators. Consistent with the limited predictive power of these interaction terms, the R^2 barely budges when we add them all to the model in Column (4), rising from 0.179 to 0.188, or by just 5 percent.

As an alternative metric for the success (or lack of success) of our candidate moderators at capturing systematic treatment effect variation, 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 Eq. (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}. Appendix Table A11 shows the results. This metric confirms the message from Table 5: the new FH lower bounds on the impact standard deviation equal 1.05 SDs for the full-cost program and 0.64 from the reduced-cost program. In both cases these represent only slight decreases from the original values in Table 2.

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).[53] ML has several advantages for the examination of systematic treatment effect heterogeneity relative to the conventional approach we applied in Section 7.2.[54] First, it allows for an exhaustive model search across a space defined by the researcher. Second, it reduces the number "researcher degrees of freedom" (Simmons et al., 2011) by automating the model-selection process, preventing researchers from "cherry picking" results they like. Third, newer ML methods address problems related to over-fitting and post-model-selection inference.[55] 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 from 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.[56]

The literature offers two broad categories of ML techniques for systematizing treatment effect heterogeneity.[57] 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.[58]

The second ("arboreal") category comprises variants of moderator selection algorithms based on regression trees. In this context, regression trees split the sample based on the values of particular moderators according to some criterion related to the amount of treatment effect heterogeneity obtained. A sequence of repeated splits forms a (causal) 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 moderators used to perform the splits randomized among the trees, constitutes a causal forest.[59]

Direct application of particular ML methods to estimate individual treatment effect heterogeneity typically requires strong, untestable assumptions to obtain consistent estimates and valid inferences. Chernozhukov et al. (2020) (hereinafter "CDDF") develop a general framework for treatment effect heterogeneity in RCTs that avoids these statistical issues under weaker assumptions, while allowing the researcher to apply their preferred ML algorithm (or conduct a "horse race" among several algorithms). They accomplish this feat in two ways: First, they focus their statistical attention not on individual treatment effect predictions but instead on various functionals of such predictions. Second, they incorporate repeated sample splitting, and its associated variance component, into their variance estimator and inference procedures. In what follows, we apply their framework to our data.

7.3.2. Description and implementation

The CDDF procedure builds on estimates obtained via repeated random splits of the raw data into main and auxiliary samples.[60] CDDF use 100 equal splits; to reduce computation time in our larger sample, we use 50 splits instead.[61] In each split, the CDDF procedure applies the researcher's preferred ML method to the auxiliary sample to estimate a control group conditional mean function $B(X) = E[Y_0|X]$ and a "proxy predictor" $S(X) = E[Y_1 - Y_0|X]$ of the population Conditional Average Treatment Effect (CATE), $s_0(X)$. Estimation of the CATE requires strong assumptions to obtain consistent estimates and valid inference, so the CDDF method focuses on estimating key *features* of the CATE instead of the CATE itself. We present two of these features: the Best Linear Predictor (BLP) of the CATEs and the Sorted Group Average Treatment Effects (GATES).[62]

---

53 Old people think "ML" denotes "Maximum Likelihood". 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. 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). 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.

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

55 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.

56 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.

57 James et al. (2017) provide an excellent textbook treatment of ML methods.

58 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. See, e.g., Chen et al. (2017), Imai and Ratkovic (2013), Knaus et al. (2020), and Tian et al. (2014), for more detail on the LASSO and empirical applications in different substantive domains.

59 See, e.g., Wager and Athey (2018), Davis and Heller (2017), Foster et al. (2011), Green and Kern (2012), Hill (2011), and Hill and Su (2013).

60 Their framework readily generalizes to stratified RCTs; we omit the associated details for simplicity.

61 They do not provide any formal guidance on how to choose the number of splits.

62 SGATEs seems more correct to us, but we nonetheless follow CDDF (2020) in acronymic misbehavior.

Table 6
BLPs of CATEs for full- and reduced-cost NULP interventions (elastic net).

	ATE (1)	HET (2)
Full-Cost Program	1.335	1.133
	(1.186,1.492)	(0.887,1.408)
Reduced-Cost Program	0.836	1.182
	(0.699,0.975)	(0.848,1.521)

Notes: Best Linear Predictors (BLPs) are estimated using the approach of CDDF (2020). The table presents median values over 50 random splits of the sample; 90% confidence intervals in parentheses. We present the elastic net results as those provide the highest performance out of the four methods (see Appendix Table A12).

Estimation of the linear model,

$$Y = \alpha_0 + \alpha_1 \tilde{B}(X) + \beta_1(D - E(D)) + \beta_2(D - E(D))(\tilde{S}(X) - E(\tilde{S}(X))) + u \tag{9}$$

using the main sample yields the coefficients of the BLP of $s_0(X)$. In equation Eq. (9), $\tilde{B}(X)$ and $\tilde{S}(X)$ denote predicted values based on the ML estimates from the auxiliary data. The BLP equals

$$BLP[s_0(X)|\tilde{S}(X)] = \hat{\beta}_1 + \hat{\beta}_2(\tilde{S}(X) - E(\tilde{S}(X))) \tag{10}$$

where $\hat{\beta}_1$ corresponds to $E(s_0(X))$ (the average treatment effect, or ATE) and $\hat{\beta}_2$ corresponds to $Cov(s_0(X), S(X))/Var(S(X))$ (the loading of the treatment effect heterogeneity on the proxy predictor ($S(X)$), or HET). In a common effect world $\beta_2 = 0$, as it would in a heterogeneous treatment effects world in which the available Xs lack the relevant moderators.

The BLPs then allow the construction of the GATES. We follow CDDF in defining groups based on quintiles of the proxy predictors. An implicit bias–variance tradeoff underlies the choice of the number of groups; quintiles work fine given our methodological aims. Unlike conventional methods, ML methods do not automatically overfit the data when given as many variables as observations. This avoids the problem of perfectly predicting the school-level variation in test scores described in Section 7.2. We therefore include all the school-level moderators in our main estimates.[63]

Following CDDF, we compare four standard ML methods: elastic net, neural network, random forest, and boosted trees.[64] The elastic net fits in our LASSO-related category, while the forest and the trees fit in our arboreal category.[65] Appendix Table A12 compares the fit-based performance measures for the four machine learning methods. The elastic net performs best on both measures for both treatments. Thus, we focus on the elastic net results for the rest of our analysis, although the results do not systematically change when considering alternative methods.

7.3.3. Findings

Table 6 presents the coefficients from the BLP of the CATE based on the machine learning proxies ($S(X)$). We present estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$, which correspond to the (ATE) and heterogeneity loading (HET) parameters, respectively. The ATE estimates in Column (1) indicate that the full- and the reduced-cost versions of the program increase test scores by 1.34 and 0.84 SDs, respectively, very close, as expected, to the estimates in Table 1. We additionally find strong heterogeneity in treatment effects, as indicated by the statistically significant and substantively meaningful heterogeneity estimates. These results corroborate our previous findings about the heterogeneity of the impacts.

We also estimate GATES by quintiles of the proxy predictor, $S(X)$. Figs. 3(a) and 3(b) present the estimates for the full- and reduced-cost versions, respectively. We find positive point estimates across all quintiles for both versions of the program, ranging from 0.47 to 2.14 SDs for the full-cost version and 0.09 to 1.31 SDs for the reduced-cost version. In Appendix Table A13 we present a formal test of the difference between these point estimates; there is a significant difference of 1.78 SDs for the full-cost treatment and of 1.22 SDs for the reduced-cost version when comparing the most- and least-affected students. These differences are substantively large, but small relative to the ranges implied by our FH lower bounds in Table 2. This leads to contrasting findings: the ML estimates imply that only the reduced-cost version of the program could have negative results (see Appendix Table A13), while the FH bounds are consistent with negative treatment effects for both versions of the program.

Appendix Table A14 presents the estimated FH bounds after removing systematic variation using the machine-learning algorithm. We subtract the BLP of the CATE from the EGRA test score index and re-estimate the FH bounds on this

63 We continue to omit moderators with severe missing data problems, recognizing the tradeoff between having more variables vs. smaller sample size.

64 For each ML method, and for each sample split, we choose the tuning parameters based on the estimates of the mean squared error of a two-fold cross-validation. Tuning and training the ML methods are done in the auxiliary sample. Due to computational time, CDDF do not use cross-validation when using random forest, and instead use the default tuning parameter. In our case, we use 50 sample splits (half of the splits in CDDF) which allow us to implement the two-fold cross-validation even when using random forest.

65 The neural network method does not fit well in either category, but is closer in spirit to the LASSO-related category.

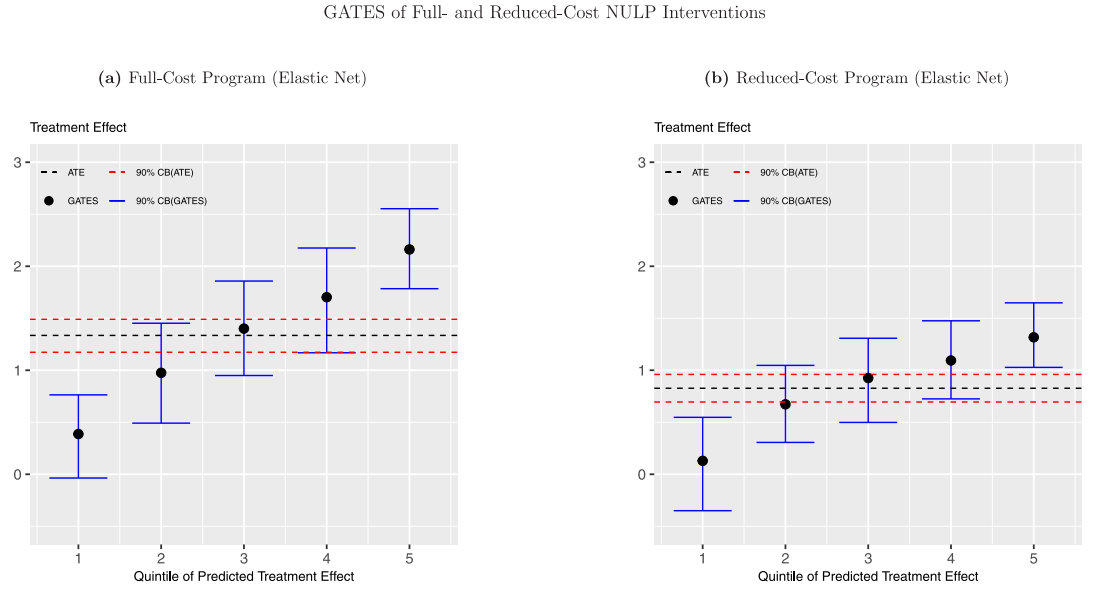

Fig. 3. GATES of Full- and Reduced-Cost NULP Interventions. *Notes*: Sorted Group Average Treatment Effects (GATES) are estimated using the approach of Chernozhukov et al. (2020) as described in Section 7.3.2. Point estimates by ML proxy quintile and joint uniform 90 percent confidence intervals are estimated based on 50 random splits of the sample.

Table 7
Fréchet–Höffding bounds removing systematic variation.

	Full-cost Program		Reduced-cost Program	
	Rank preservation (1)	Rank inversion (2)	Rank preservation (3)	Rank inversion (4)
Unadjusted FH Bounds	1.066	2.615	0.642	2.218
	[1.019,1.107]	[2.586,2.645]	[0.609,0.686]	[2.195,2.240]
Removing Systematic Variation:				
Conventional Method	1.052	2.612	0.636	2.223
	[1.006,1.092]	[2.586,2.640]	[0.602,0.678]	[2.201,2.243]
Machine Learning	0.985	2.544	0.609	2.191
	[0.935,1.025]	[2.515,2.570]	[0.575,0.652]	[2.169,2.212]

Notes: Columns (1) and (3) show statistics estimated using the Fréchet–Höffding lower-bound distribution from Eq. (2), while Columns (2) and (4) use the upper-bound distribution. Rows 2 and 3 of the table show the FH bounds after removing systematic variation using the conventional method (as described in Section 7.2) and the machine-learning method (as described in Section 7.3). All estimates use the main analysis sample. Bias-corrected confidence intervals, bootstrapped using 1000 replications, in brackets.

vector.[66] For ease of comparison, we present the unadjusted FH bounds along with the bounds obtained using both the conventional as well as the ML method to remove systematic variation in Table 7. As with the conventional approach, removing systematic variation using ML leads to only a very modest change from the unadjusted bounds: the lower-bound impact standard deviation falls by just 8 percent for the full-cost version of the program.

7.4. Could "better data help a lot"?

The overarching conclusion from the analyses in the two preceding sections is that we do a very poor job 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 ways to view this finding. The pessimistic view 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 view sees it as a pointed reminder that researchers have not really pushed very hard on the theory or measurement of effect moderation. At the margin, more effort on

66 Appendix A2 presents the FH bounds estimates excluding school-level moderators. The best method continues to be the elastic net (Appendix Table A15); the FH lower bound is nearly unchanged for the full-cost program and about four percent lower for the reduced-cost program (Appendix Table A16). The estimates in our IZA working paper (Buhl-Wiggers et al., 2020), which used the LASSO-based method of Knaus et al. (2020), yield a lower bound of the impact SD of 1.02 for the full-cost program, and 0.65 for the reduced-cost version.

theory and data collection might have a substantially higher payoff than the same amount of time and effort devoted to tweaking the ML algorithm du jour.

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. For example, 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.

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. Similarly, we have no 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.

8. Conclusion

Using data from a randomized evaluation of a highly effective literacy program in 128 primary schools in northern Uganda, we resoundingly reject the null hypothesis of equal student-level treatment effects. The full-cost program's average gain of 1.40 SDs masks the fact that, assuming normally distributed treatment effects, at least 29 percent of students experience a gain of more than 2 SDs, while the program makes more than 10 percent of students worse off. For the full-cost version of the program, the FH lower bound on the impact standard deviation exceeds 1.0 SDs of our endline Leblango reading test score index. The variation in gains within this program exceeds the difference in the mean effects across the two versions implemented in our study. There is also 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.

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 (2021) concentrates any possible negative average effects at the upper end of the outcome distribution but otherwise delivers disappointingly 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 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. 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 help address the "learning crisis", in which many students end up learning nothing even after years of attending school (World Bank, 2018). The "success" of a learning intervention – such as the highly effective intervention we study in this paper – 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 suggest 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 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 little in the way of systematic treatment effect heterogeneity, even when exploited by state-of-the-art ML algorithms. 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 heretofore unexamined potential moderators. Better data could yield higher returns at the margin than further refinements to existing ML methods.

[67] This idea sits at the core of the "Teaching at the Right Level" program in Banerjee et al. (2017); the US-based Response to Intervention method also targets interventions by student performance levels; see, e.g. Mesmer and Mesmer (2008).

[68] Duflo et al. (2011) provide an example of the effectiveness of tracking in Kenya.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.12.010.

References

Abadie, Alberto, Angrist, Joshua, Imbens, Guido, 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70 (1), 91–117. http://dx.doi.org/10.1111/1468-0262.00270.

Altinyelken, Hulya Kosar, 2010. Curriculum change in Uganda: Teacher perspectives on the new thematic curriculum. Int. J. Educ. Dev. 30 (2), 151–161. http://dx.doi.org/10.1016/j.ijedudev.2009.03.004.

Angrist, Joshua D., Lavy, Victor, 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. 114 (2), 533–575. http://dx.doi.org/10.1162/003355399556061.

Athey, Susan, 2019. The impact of machine learning on economics. In: Agrawal, Ajay, Gans, Joshua, Goldfarb, Avi (Eds.), The Economics of Artificial Intelligence: an Agenda. University of Chicago Press, Chicago, pp. 507–552. http://dx.doi.org/10.7208/9780226613475-023.

Banerjee, Abhijit, Banerji, Rukmini, Berry, James, Duflo, Esther, Kannan, Harini, Mukerji, Shobhini, Shotland, Marc, Walton, Michael, 2017. From proof of concept to scalable policies: Challenges and solutions, with an application. J. Econ. Perspect. 31 (4), 73–102. http://dx.doi.org/10.1257/jep.31.4.73.

Berger, Mark C., Black, Dan, Smith, Jeffrey A., 2001. Evaluating profiling as a means of allocating government services. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. In: ZEW Economic Studies (Publication Series of the Centre for European Economic Research (ZEW), Mannheim, Germany), vol. 13, Physica, Heidelberg, pp. 59–84. http://dx.doi.org/10.1007/978-3-642-57615-7_4.

Bhattacharya, Jay, Shaikh, Azeem M., Vytlacil, Edward, 2008. Treatment effect bounds under monotonicity assumptions: An application to Swan-Ganz catheterization. Amer. Econ. Rev. 98 (2), 351–356. http://dx.doi.org/10.1257/aer.98.2.351.

Bitler, Marianne P., Gelbach, Jonah B., Hoynes, Hilary W., 2005. Distributional Impacts of the Self-Sufficiency Project. Working Paper 11626, National Bureau of Economic Research, http://dx.doi.org/10.3386/w11626.

Bitler, Marianne P., Gelbach, Jonah B., Hoynes, Hilary W., 2006. What mean impacts miss: Distributional effects of welfare reform experiments. Amer. Econ. Rev. 96 (4), 988–1012. http://dx.doi.org/10.1257/aer.96.4.988.

Bitler, Marianne P., Gelbach, Jonah B., Hoynes, Hilary W., 2017. Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? Evidence from a social experiment. Rev. Econ. Stat. http://dx.doi.org/10.1162/REST_a_00662.

Bitler, Marianne P., Hoynes, Hilary W., Domina, Thurston, 2014. Experimental Evidence on Distributional Effects of Head Start. Working Paper 20434, National Bureau of Economic Research, http://dx.doi.org/10.3386/w20434.

Black, Dan A., Smith, Jeffrey A., 2004. How robust is the evidence on the effects of college quality? Evidence from matching. J. Econometrics vol. 121 (1), 99–124. http://dx.doi.org/10.1016/j.jeconom.2003.10.006, Higher education (Annals issue).

Bold, Tessa, Filmer, Deon, Martin, Gayle, Molina, Ezequiel, Stacy, Brian, Rockmore, Christophe, Svensson, Jakob, Wane, Waly, 2017. Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa. J. Econ. Perspect. 31 (4), 185–204. http://dx.doi.org/10.1257/jep.31.4.185.

Boone, Peter, Fazzio, Ila, Jandhyala, Kameshwari, Jayanty, Chitra, Jayanty, Gangadhar, Johnson, Simon, Ramachandran, Vimala, Silva, Filipa, Zhan, Zhaoguo, 2014. 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). In: African Successes, Volume II: Human Capital. University of Chicago Press, pp. 255–280. http://dx.doi.org/10.7208/chicago/9780226316192.001.0001.

Bruhn, Miriam, McKenzie, David, 2009. In Pursuit of balance: Randomization in practice in development field experiments. Am. Econ. J.: Appl. Econ. 1 (4), 200–232. http://dx.doi.org/10.1257/app.1.4.200.

Brunette, Tracy, Piper, Benjamin, Jordan, Rachel, King, Simon, Nabacwa, Rehemah, 2019. The impact of mother tongue reading instruction in twelve ugandan languages and the role of language complexity, socioeconomic factors, and program implementation. Comp. Educ. Rev. 63 (4), 591–612. http://dx.doi.org/10.1086/705426.

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 Paper 13598, Institute for the Study of Labor (IZA), Bonn, Germany.

Buhl-Wiggers, Julie, Kerwin, Jason, Muñoz-Morales, Juan, Smith, Jeffrey, Thornton, Rebecca, 2022. Replication data for SCLB. Harvard Dataverse, http://dx.doi.org/10.7910/DVN/WUMJOJ.

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

Buhl-Wiggers, Julie, Kerwin, Jason, Smith, Jeffrey, Thornton, Rebecca, 2018b. Teacher Effectiveness in Africa: Longitudinal and Causal Estimates. IGC Working Paper S-89238-UGA-1, International Growth Centre.

Cambanis, Stamatis, Simons, Gordon, Stout, William, 1976. Inequalities for E(k(x,y)) when the marginals are fixed. Z. Wahrscheinlichkeit. Und Verwandte Gebiete 36, 285–294. http://dx.doi.org/10.1007/BF00532695.

Chen, Shuai, Tian, Lu, Cai, Tianxi, Yu, Menggang, 2017. A general statistical framework for subgroup identification and comparative treatment scoring. Biometrics 73 (4), 1199–1209. http://dx.doi.org/10.1111/biom.12676.

Chernozhukov, Victor, Demirer, Mert, Duflo, Esther, Fernández-Val, Iván, 2020. Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. Working Paper 24678, National Bureau of Economic Research, http://dx.doi.org/10.3386/w24678.

Chernozhukov, Victor, Fernández-Val, Iván, 2005. Subsampling inference on quantile regression processes. ankhyā: Indian J. Stat. (2003-2007) 67 (2), 253–276.

Chung, EunYi, Olivares, Mauricio, 2021. Permutation test for heterogeneous treatment effects with a nuisance parameter. J. Econometrics 225 (2), 148–174. http://dx.doi.org/10.1016/j.jeconom.2020.09.015, Themed Issue: Treatment Effect 1.

Cilliers, Jacobus, Fleisch, Brahm, Prinsloo, Cas, Taylor, Stephen, 2020. How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. J. Hum. Resour. 55, 926–962. http://dx.doi.org/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. Rev. Educ. Res. 87 (5), 863–898. http://dx.doi.org/10.3102/0034654317712025.

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

Cummins, Joseph R., 2017. Heterogeneous treatment effects in the low track: Revisiting the Kenyan Primary School experiment. Econ. Educ. Rev. 56, 40–51. http://dx.doi.org/10.1016/j.econedurev.2016.11.006.

Davis, Jonathan M.V., Heller, Sara B., 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. Amer. Econ. Rev. 107 (5), 546–550. http://dx.doi.org/10.1257/aer.p20171000.

Delavallade, Clara, Griffith, Alan, Thornton, Rebecca, 2021. Effects of a multi-faceted education program on enrollment, learning and gender equity: Evidence from India. World Bank Econ. Rev. http://dx.doi.org/10.1093/wber/lhaa025.

Djebbari, Habiba, Smith, Jeffrey, 2008. Heterogeneous impacts in progresa. J. Econometrics 145 (1), 64–80. http://dx.doi.org/10.1016/j.jeconom.2008.05.012.

Doksum, Kjell, 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. Ann. Statist. 267–277. http://dx.doi.org/10.1214/aos/1176342662.

Dong, Yingying, Shen, Shu, 2018. Testing for rank invariance or similarity in program evaluation. Rev. Econ. Stat. 100 (1), 78–85. http://dx.doi.org/10.1162/REST_a_00686.

Duflo, Esther, Dupas, Pascaline, Kremer, Michael, 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. Amer. Econ. Rev. 101 (5), 1739–1774. http://dx.doi.org/10.1257/aer.101.5.1739.

Evans, David K., Popova, Anna, 2016. What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. World Bank Res. Observ. 31 (2), 242–270. http://dx.doi.org/10.1093/wbro/lkw004.

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

Evans, David K., Yuan, Fei, 2019. What We Learn About Girls' Education from Interventions that Do Not Focus on Girls. Policy Research Working Papers, The World Bank, http://dx.doi.org/10.1596/1813-9450-8944.

Fan, Yanqin, Park, Sang Soo, 2010. Sharp bounds on the distribution of treatment effects and their statistical inference. Econom. Theory 26 (3), 931–951. http://dx.doi.org/10.1017/S0266466609990168.

Foster, Jared C., Taylor, Jeremy M.G., Ruberg, Stephen J., 2011. Subgroup identification from randomized clinical trial data. Stat. Med. 30 (24), 2867–2880. http://dx.doi.org/10.1002/sim.4322.

Frandsen, Brigham R., Lefgren, Lars J., 2017. Testing rank similarity. Rev. Econ. Stat. 100 (1), 86–91. http://dx.doi.org/10.1162/REST_a_00675.

Frandsen, Brigham R., Lefgren, Lars J., 2021. Partial identification of the distribution of treatment effects with an application to the knowledge is power program (KIPP). Quant. Econ. 12 (1), 143–171. http://dx.doi.org/10.3982/QE1273.

Fréchet, M., 1951. Les tableaux de corrélation dont les marges sont données. Ann. L'Univer. Lyon. Sec. A: Sci., Math. Et Astron. 14, 53–77. http://dx.doi.org/10.2307/1401846.

Freedman, David A., 2008. On regression adjustments to experimental data. Adv. Appl. Math. 40 (2), 180–193. http://dx.doi.org/10.1016/j.aam.2006.12.003.

Ganimian, Alejandro J., Murnane, Richard J., 2014. Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Impact Evaluations. Working Paper 20284, National Bureau of Economic Research, http://dx.doi.org/10.3386/w20284.

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/.

Gilligan, Daniel O., Karachiwalla, Naureen, Kasirye, Ibrahim, Lucas, Adrienne M., Neal, Derek, 2019. Educator incentives and educational triage in rural primary schools. J. Hum. Resour. 1118. http://dx.doi.org/10.3368/jhr.57.1.1118-9871R2.

Glewwe, Paul W., Hanushek, Eric A., Humpage, Sarah D., Ravina, Renato, 2013. School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In: Glewwe, Paul W. (Ed.), Education Policy in Developing Countries. University of Chicago Press, Chicago and London, http://dx.doi.org/10.7208/chicago/9780226078854.003.0002.

Glewwe, Paul, Kremer, Michael, Moulin, Sylvie, 2009. Many children left behind? Textbooks and test scores in Kenya. Am. Econ. J.: Appl. Econ. 1 (1), 112–135. http://dx.doi.org/10.1257/app.1.1.112.

Glewwe, Paul, Muralidharan, Karthik, 2016. Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In: Hanushek, Eric A., Machin, Stephen, Woessmann, Ludger (Eds.), Handbook of the Economics of Education, Vol. 5. Elsevier, pp. 653–743. http://dx.doi.org/10.1016/B978-0-444-63459-7.00010-5.

Gove, Amber, Brunette, Tracy, Bulat, Jennae, Carrol, Bidemi, Henny, Catherine, Macon, Wykia, Nderu, Evangeline, Sitabkhan, Yasmin, 2017. Assessing the impact of early learning programs in Africa. New Dir. Child Adolesc. Dev. 2017 (158), 25–41. http://dx.doi.org/10.1002/cad.20224.

Green, Donald P., Kern, Holger L., 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public Opin. Q. 76 (3), 491–511. http://dx.doi.org/10.1093/poq/nfs036.

Guggenberger, Patrik, 2010. The impact of a Hausman pretest on the asymptotic size of a hypothesis test. Econom. Theory 26 (2), 369–382. http://dx.doi.org/10.1017/S0266466609100026.

Hanushek, Eric A., 1992. The trade-off between child quantity and quality. J. Polit. Econ. 100 (1), 84–117. http://dx.doi.org/10.1086/261808.

Hanushek, Eric A., Rivkin, Steven G., 2010. Generalizations about using value-added measures of teacher quality. Amer. Econ. Rev. 100 (2), 267–271. http://dx.doi.org/10.1257/aer.100.2.267.

Heckman, James, Ichimura, Hidehiko, Smith, Jeffrey, Todd, Petra, 1998. Characterizing selection bias using experimental data. Econometrica 66 (5), 1017–1098. http://dx.doi.org/10.2307/2999630.

Heckman, James J., Kautz, Tim, 2012. Hard evidence on soft skills. Labour Econ. 19 (4), 451–464. http://dx.doi.org/10.1016/j.labeco.2012.05.014, European Association of Labour Economists 23rd annual conference, Paphos, Cyprus, 22-24th September 2011.

Heckman, J.J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. Rev. Econom. Stud. 64 (4), 487–535. http://dx.doi.org/10.2307/2971729.

Heckman, James J., Urzua, Sergio, Vytlacil, Edward, 2006. Understanding instrumental variables in models with essential heterogeneity. Rev. Econ. Stat. 88 (3), 58. http://dx.doi.org/10.1162/rest.88.3.389.

Hill, Jennifer L., 2011. Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Statist. 20 (1), 217–240. http://dx.doi.org/10.1198/jcgs.2010.08162.

Hill, Jennifer, Su, Yu-Sung, 2013. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. Ann. Appl. Stat. 1386–1420. http://dx.doi.org/10.1214/13-AOAS630.

Höffding, W., 1940. Masstabinvariante korrelationsmasse für diskontinuierliche verteilungen. Arkiv Mat. Wirtsch. Sozialforschung 7, 49–70.

Hotz, V. Joseph, Imbens, Guido W., Mortimer, Julie H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. J. Econometrics 125 (1–2), 241–270. http://dx.doi.org/10.1016/j.jeconom.2004.04.009.

Imai, Kosuke, Ratkovic, Marc, 2013. Estimating treatment effect heterogeneity in randomized program evaluation. Ann. Appl. Stat. 7 (1), 443–470. http://dx.doi.org/10.1214/12-AOAS593.

Jackson, Kirabo, Makarin, Alexey, 2018. Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment. Am. Econ. J.: Econ. Policy 10 (3), 226–254. http://dx.doi.org/10.1257/pol.20170211.

James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert, 2017. An Introduction to Statistical Learning, eighth ed. Springer, ISBN: 978-1-4614-7137-0.

Jiménez, Juan E., Gove, Amber, Crouch, Luis, 2014. Internal structure and standardized scores of the spanish adaptation of the EGRA (early grade reading assessment) for early reading assessment. Psicothema (26.4), 531–537. http://dx.doi.org/10.7334/psicothema2014.93.

Kerwin, Jason T., Thornton, Rebecca L., 2021. Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. Rev. Econ. Stat. 103 (2), 251–264. http://dx.doi.org/10.1162/rest_a_00911.

Knaus, Michael C., Lechner, Michael, Strittmatter, Anthony, 2020. Heterogeneous employment effects of job search programmes: A machine learning approach. J. Hum. Resour. http://dx.doi.org/10.3368/jhr.57.2.0718-9615R1.

Koenker, Roger, Bassett, Gilbert, 1978. Regression quantiles. Econometrica 46 (1), 33–50. http://dx.doi.org/10.2307/1913643.
Koenker, Roger, Bilias, Yannis, 2001. Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments. Empir. Econ. 26 (1), 199–220. http://dx.doi.org/10.1007/s001810000057.
Kremer, Michael, Brannen, Conner, Glennerster, Rachel, 2013. The challenge of education and learning in the developing world. Science 340 (6130), 297–300. http://dx.doi.org/10.1126/science.1235350.
Krishnaratne, Shari, White, Howard, Carpenter, Ella, 2013. Quality Education for All Children? What Works in Education in Developing Countries. Working Paper 20, International Initiative for Impact Evaluation (3ie), New Delhi.
Lee, Soohyung, Shaikh, Azeem M., 2014. Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of progresa on school enrollment: Heterogeneous treatment effects. J. Appl. Econometrics 29 (4), 612–626. http://dx.doi.org/10.1002/jae.2327.
Lehmann, Erich Leo, D'Abrera, Howard J., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.
Lim, Jaegeum, Meer, Jonathan, 2017. The impact of teacher–student gender matches: Random assignment evidence from South Korea. J. Hum. Resour. 52 (4), 979–997. http://dx.doi.org/10.3368/jhr.52.4.1215-7585R1.
Lin, Winston, 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's Critique. Ann. Appl. Stat. 7 (1), 295–318. http://dx.doi.org/10.1214/12-AOAS583.
Linhart, H., Zucchini, W., 1986. Model Selection. Wiley, New York, ISBN: 978-0-471-83722-0.
McEwan, Patrick J., 2015. Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. Rev. Educ. Res. 85 (3), 353–394. http://dx.doi.org/10.3102/0034654314553127.
Mesmer, Eric M., Mesmer, Heidi Anne E., 2008. Response to intervention (RTI): What teachers of reading need to know. Reading Teacher 62 (4), 280–290. http://dx.doi.org/10.1598/RT.62.4.1.
Moshoeshoe, Ramaele, 2015. Average and Heterogeneous Effects of Class Size on Educational Achievement in Lesotho. Working Papers 496, Economic Research Southern Africa.
Piper, Benjamin, 2010. Uganda Early Grade Reading Assessment Findings Report: Literacy Acquisition and Mother Tongue. Research Triangle Institute.
Piper, Benjamin, Zuilkowski, Stephanie Simmons, Kwayumba, Dunston, Oyanga, Arbogast, 2018. Examining the secondary effects of mother-tongue literacy instruction in Kenya: Impacts on student learning in english, kiswahili, and mathematics. Int. J. Educ. Dev. 59, 110–127. http://dx.doi.org/10.1016/j.ijedudev.2017.10.002.
Piper, Benjamin, Zuilkowski, Stephanie S., Ong'ele, Salome, 2016. Implementing mother tongue instruction in the real world: Results from a medium-scale randomized controlled trial in Kenya. Comp. Educ. Rev. 60 (4), 776–807. http://dx.doi.org/10.1086/688493.
Powell, David, 2020. Quantile treatment effects in the presence of covariates. Rev. Econ. Stat. 102 (5), 994–1005. http://dx.doi.org/10.1162/rest_a_00858.
Pritchett, Lant, Beatty, Amanda, 2015. Slow down, you're going too fast: Matching curricula to student skill levels. Int. J. Educ. Dev. 40, 276–288. http://dx.doi.org/10.1016/j.ijedudev.2014.11.013.
RTI International, 2019. Early grade reading assessment toolkit.
Rudalevige, Andrew, 2003. The politics of no child left behind. Educ. Next 3 (4), 63–69.
Simmons, Joseph P., Nelson, Leif D., Simonsohn, Uri, 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366. http://dx.doi.org/10.1177/0956797611417632.
Ssentanda, Medadi Erisa, Huddlestone, Kate, Southwood, Frenette, 2016. The politics of mother tongue education: The case of Uganda. Per Linguam 32 (3), 60–78. http://dx.doi.org/10.5785/32-3-689.
Tchen, André H., 1980. Inequalities for distributions with given marginals. Ann. Probab. 8 (4), 814–827.
Tian, Lu, Alizadeh, Ash A., Gentles, Andrew J., Tibshirani, Robert, 2014. A simple method for estimating interactions between a treatment and a large number of covariates. J. Amer. Statist. Assoc. 109 (508), 1517–1532. http://dx.doi.org/10.1080/01621459.2014.951443.
Todd, Petra E., Wolpin, Kenneth I., 2003. On the specification and estimation of the production function for cognitive achievement. Econ. J. 113 (485), F3–F33. http://dx.doi.org/10.1111/1468-0297.00097.
Uganda Bureau of Statistics, 2017. The national population and housing census 2014 – education in the thematic report series.
UNESCO, 2017. A Guide for Ensuring Inclusion and Equity in Education. United Nations Educational, Scientific and Cultural Organization, Paris, France, ISBN 978-92-3-100222-9.
Uwezo, 2016. Are Our Children Learning? Uwezo Uganda Eighth Learning Assessment Report. Twaweza East Africa, Kampala.
Wager, Stefan, Athey, Susan, 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Amer. Statist. Assoc. 113 (523), 1228–1242. http://dx.doi.org/10.1080/01621459.2017.1319839.
Weiss, Michael J., Bloom, Howard S., Brock, Thomas, 2014. A conceptual framework for studying the sources of variation in program effects. J. Policy Anal. Manag. 33 (3), 778–808. http://dx.doi.org/10.1002/pam.21760.
Williamson, Robert C., Downs, Tom, 1990. Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. Internat. J. Approx. Reason. 4 (2), 89–158. http://dx.doi.org/10.1016/0888-613X(90)90022-T.
World Bank, 2018. World Development Report 2018: Learning to Realize Education's Promise. Tech. Rep., World Bank, Washington, DC, http://dx.doi.org/10.1596/978-1-4648-1096-1.